CMMR 2019

MARSEILLE

Perception Representations
Image Sound
Music

14 - 18 october
14 th International
Symposium on Computer Music
Multidisciplinary Research

PROCEEDINGS

Proceedings of the

14th International Symposium on Computer Music Multidisciplinary Research

14 – 18th October, 2019
Marseille, France

Organized by

The Laboratory PRISM
"Perception, Representations, Image, Sound, Music"
Marseille, France

in collaboration with

GMEM "Groupe de Musique Expérimentale de Marseille"
and
n + n corsino

Aix*Marseille université
initiative d'excellence

PRISM

Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes

gmem
CENTRE NATIONAL DE CRÉATION MUSICALE

n + n corsino

Published by

The Laboratory PRISM
"Perception, Representations, Image, Sound, Music"
31, chemin Joseph Aiguier
CS 70071
13402 Marseille Cedex 09 - France

October, 2019

Proceedings Editors: M. Aramaki, O. Derrien,
R. Kronland-Martinet, S. Ystad

ISBN 979-10-97-498-01-6
Les éditions de PRISM

Welcome to CMMR 2019!

We are happy to welcome you to the 14th edition of CMMR in Marseille. This is the second CMMR event that takes place in Marseille, but in a slightly different context than in 2013, since the present edition is organized by the new interdisciplinary art-science laboratory, PRISM (Perception, Representations, Image, Sound, Music), which very much reflects the spirit of the CMMR conference cycle. PRISM hosts researchers within a large variety of fields, spanning from physics and signal processing, art and aesthetic sciences to medicine and neuroscience that all have a common interest in the perception and representation of image, sound and music. The scientific challenge of PRISM is to reveal how the audible, the visible and their interactions generate new forms of sensitive and/or formal representations of the contemporary world.

CMMR2019 will be the occasion to celebrate the creation of the PRISM and at the same time honor one of its co-founders, researcher, composer and computer music pioneer Jean-Claude Risset who sadly passed away in November 2016, only two months before the laboratory was officially acknowledged. A scientific session followed by an evening concert will be dedicated to him on the first day of the conference.

From the first announcement of the CMMR2019 we received a large response from both scientists and artists who wanted to participate in the conference, either by organizing special sessions, presenting demos or installations or proposing workshops and concerts. Among the 15 scientific sessions that will take place during the conference, eight special sessions that deal with various subjects from sound design, immersive media and mobile devices to music and deafness, embodied musical interaction and phenomenology of the conscious experience are scheduled. We are also lucky to have three internationally renowned keynote speakers with us during this edition: John Chowning, Professor Emeritus at Stanford University who will talk about his friend and colleague Jean-Claude Risset, Geoffroy Peeters, Professor at Télécom ParisTech who will talk about past and present research within Music Information Research and Josh McDermott, Associate Professor in the Department of Brain and Cognitive Sciences at MIT who will present classic and recent approaches to auditory scene analysis.

The artistic program that has been elaborated in collaboration with "n+n corsino" and GMEM includes a tribute concert to Jean-Claude Risset, scheduled on Monday evening, a virtual/augmented concert on Tuesday evening and a contemporary music concert on Wednesday evening. During the last evening, an interactive music concert will take place under the direction of Christophe Héral. Sound installations and a videomusic presentation are also scheduled during the conference.

Finally, in addition to the scientific paper, poster and demo sessions and the artistic program, five satellite workshops are programmed right after the conference on Friday October 18th.

We hope that CMMR2019 will be an unforgettable event for all of you, and wish you a pleasant stay in Marseille.

R. Kronland-Martinet, S. Ystad and M. Aramaki
The CMMR2019 symposium chairs

Organization

The 14th International Symposium on Computer Music Multidisciplinary Research CMMR2019 "Perception, Representations, Image, Sound, Music" is organized by the laboratory PRISM (CNRS-AMU, UMR 7061, France), GMEM and n+n corsino.

Symposium Chairs

Richard Kronland-Martinet (PRISM, AMU-CNRS, France)
Sølvi Ystad (PRISM, AMU-CNRS, France)
Mitsuko Aramaki (PRISM, AMU-CNRS, France)

Proceedings Chair

Olivier Derrien (Univ. Toulon, PRISM, France)

Paper Chairs

Mitsuko Aramaki (PRISM, AMU-CNRS, France)
Ivan Magrin-Chagnolleau (PRISM, AMU-CNRS, France)

Programme Chairs

Richard Kronland-Martinet (PRISM, AMU-CNRS, France)
Sølvi Ystad (PRISM, AMU-CNRS, France)

Artistic Programme Chairs

Norbert Corsino (n+n corsino, France)
Jacques Sapiéga (PRISM, AMU-CNRS, France)
Christian Sébille (GMEM, France)

Workshop Chair

Mathieu Barthet (QMUL, United Kingdom)

Demo Chair

Adrien Vidal (PRISM, AMU-CNRS, France)

Poster Chair

Samuel Poirot (PRISM, AMU-CNRS, France)

Sponsoring chairs

Antoine Bourachot (PRISM, AMU-CNRS, France)
Simon Fargeot (PRISM, AMU-CNRS, France)
Samuel Poirot (PRISM, AMU-CNRS, France)

Webmaster

Antoine Bourachot (PRISM, AMU-CNRS, France)

Local Organising Committee

Mitsuko Aramaki (PRISM CNRS-AMU, France)
Mathieu Barthet (QMUL, UK)
Antoine Bourachot (PRISM CNRS-AMU, France)
Olivier Derrien (PRISM CNRS-AMU, France)
Simon Fargeot (PRISM CNRS-AMU, France)
Antoine Gonot (PRISM CNRS-AMU, France)
Richard Kronland-Martinet (PRISM CNRS-AMU, France)
Claudine Le Van Phu (PRISM CNRS-AMU, France)
Samuel Poirot (PRISM CNRS-AMU, France)
Adrien Vidal (PRISM CNRS-AMU, France)
Sølvi Ystad (PRISM CNRS-AMU, France)

Paper Committee

Mitsuko Aramaki (PRISM CNRS-AMU, France)
Mathieu Barthet (QMUL, UK)
Jonathan Bell (PRISM CNRS-AMU, France)
Jonathan Berger (Stanford University, USA)
Gilberto Bernardes (University of Porto, Portugal)
Tifanie Bouchara (CNAM, France)
Sylvain Brétéché (PRISM CNRS-AMU, France)
Lionel Bringoux (ISM AMU-CNRS, France)
Marco Buongiorno Nardelli (University of North Texas, USA)
Amílcar Cardoso (University of Coimbra, Portugal)
Chris Chafe (Stanford University, USA)
Roger Dannenberg (Carnegie Mellon University, USA)
Matthew Davies (INESC TEC, Portugal)
Philippe Depalle (McGill University, Canada)

Table of Contents

Special Session - The Process of Sound Design (Tools, Methods, Productions)

Special Session - Sonic Interaction for Immersive Media - Virtual and Augmented Reality

Poster Session

Demo Papers

Modal Logic for Tonal Music

Satoshi Tojo*

Japan Advanced Institute of Science and Technology
tojo@jaist.ac.jp

Abstract. It is generally accepted that the origin of music and language is one and the same. Thus far, many syntactic theories of music have been proposed, however, all these efforts mainly concern generative syntax. Although such syntax is advantageous in constructing hierarchical tree, it is weak in representing mutual references in the tree. In this research, we propose the annotation of tree with modal logic, by which the reference from each pitch event to regions with harmonic functions are clarified. In addition, while the generative syntax constructs the tree in the top-down way, the modal interpretation gives the incremental construction according to the progression of music. Therefore, we can naturally interpret our theory as the expectation–realization model that is more familiar to our human recognition of music.

Keywords: Generative Syntax, Modal Logic, Kripke semantics

1 Introduction

What is the semantics of music? In general, the semantics is distinguished between the reference to the internal structure and that to the outer worlds. Even though we could somewhat guess what the external references mean,[1] still remains the question what is the internal semantics. Meyer [13] argued that we could define an innate meaning independent of the external references. Although there should be a big discussion more for this, we contend that we could devise a formal method to clarify the internal references.

Assuming that the origin of music and language is one and the same [18], we consider incorporating Montagovian semantics [6], as a parallelism between syntactic structure and logical forms, into music.[2]

In this work, we propose to employ the modal logic to represent internal references in music. Such modal operators as

* This work is supported by JSPS kaken 16H01744.

[1] Koelsch [12] further distinguished the three levels of the reference to the outer worlds; (i) the simple imitation of sounds by instruments, (ii) the implication of human emotions, and (iii) the artificial connection to our social behavior.

[2] We need to take care of the ambiguity of what *semantics* means. In Montagovian theory, the syntax of natural language is those written by categorial grammar and the semantics is written by logical formuale, while in mathematical logic the formal language (logic) has its own syntax and its semantics is given by set theory or by algebra.

$\Box \cdots$ For all the time points in some neighborhood region
$\Diamond \cdots$ A time point exists in any neighborhood region

denote the *accessibility* from each pitch event in music to specified regions, and thus, we can clarify the inter-region relationship in a piece.

Thus far, many linguistic approaches have been made to analyze the music structure, however, almost all these efforts concern the generative syntax, based on Context-free Grammar (CFG) by Chomsky [1, 2]. However, the non-terminal symbols in the production rules, which appear as nodes in the tree, are often vague in music; in some cases they may represent the saliency among pitch events and in other cases overlaying relations of functional regions. Also, the hierarchical construction of tree is weak in representing mutual phrasal relationship. In this work, we reinterpret those generative grammar rules in the X-bar theory [3], and show their *heads* explicitly. Then, we define the annotations of rules in logical formulae, and give a rigorous semantics by modal logic.

This reinterpretation accompanies one more significant aspect. The tree structure is constructed by production rules in the top-down way. However, when we listen to or compose music, we recognize it in the chronological order according to the progression of time. With our method, those rules would be transformed to construct a tree in the incremental way.

This paper is organized as follows. In the following section, we survey the syntactic studies in music. Next, we introduce the modal logic as formal language to represent the internal structure of music, regarding its neighborhood semantics as the inclusion and the order of time intervals, and then, we give a concrete example of analysis by modal logic. In the final section, we summarize our contribution and discuss our future tasks.

2 Syntactic Theory of Music

Thus far, many linguists and musicologists have achieved to implement the music parser, beginning from Winograd [19]. Some works were based on specific grammar theories; Tojo [17] employed Head-driven Phrase Structure Grammar (HPSG) and Steedman et al. [8] Combinatory Categorial Grammar (CCG). In recent years, furthermore, two distinguished works are shown; one is exGTTM by Hamanaka et al. [11] based on Generative Theory of Tonal Music (GTTM) [10] and the other is Generative Syntax Model (GSM) by Rohmeier [16].

2.1 Brief Introduction of GSM

We briefly introduce GSM with its abridged grammar for convenience. First we introduce the basic sets

$$\mathbb{R} = \{\, TR, SR, DR \,\} \qquad \text{(region)}$$
$$\mathbb{F} = \{\, t, s, d, tp, sp, dp, tcp \,\} \qquad \text{(function)}$$

as well as $\mathbb{K}$: a set of key names and $\mathbb{O}$: a set of chord names.

The P (piece) is the start symbol of production rules. It introduces TR (tonic region)

$$P \to TR,$$

In the next level, TR generates DR (dominant region) and SR (subdominant region), and in the further downward level they result in t (tonic), d (dominant), s (subdominant).

$$\begin{aligned} TR &\to DR\ t & \qquad TR &\to t \\ DR &\to SR\ d & \qquad DR &\to d \\ TR &\to TR\ DR & \qquad SR &\to s \end{aligned}$$

Also, each of t, d, s may result in tp (tonic parallel), tcp (tonic counter-parallel), dp (dominant parallel), sp (subdominant parallel) of Hugo Riemann [9].

$$t \to tp \mid tcp,\ s \to sp,\ d \to dp$$

where the vertical bar ('|') shows the 'or' alternatives. In addition, there are scale-dgree rules, which maps function names to degrees, *e.g.*,

$$t \to \mathrm{I},\ tp \to \mathrm{VI} \mid \mathrm{III},\ s \to \mathrm{IV},\ d \to \mathrm{V} \mid \mathrm{VII},\ \text{and so on.}$$

Furthermore, [16] employed the rules of secondary dominant, and that of modulation.

However, the formalism of GSM is partially inconvenient since the interpretation of

$$XR \to XR\ XR\ (XR \in \mathbb{R})$$

is ambiguous as to which is the head. Rather, we prefer to rigorously define heads, distinguishing region rules from harmonic function rules.

2.2 *X*-barred GSM

Here, we introduce the notion of *head*, employing the X-bar theory [3].[3] A binary production in Chomsky Normal Form (CNF) should be rewritten with the parent category X' and the head daughter category X, as follows.

$$\begin{cases} X' \to Spec\ X \\ X' \to X\ \ Comp \end{cases}$$

where *Spec* stands for specifier and *Comp* for complement.[4] When X in the right-hand side already includes a prime ($'$), X' in the left-hand side becomes doubly primed ($''$) recursively.

[3] The latest Chomskian school has abandoned X-bar theory, and instead they explain every syntactic phenomena only by *merge* and *recursion* [4].

[4] Note that notion of head resides also in Combinatory Categorial Grammar (CCG), as $X \to X/Z\ Z$ implies that X/Z is the principle constituent of X.

Since we do not use those degree rules, our simplified GSM rules becomes as follows, where $\mathbb{F} = \{tnc, dom, sub\}$.

$$
\begin{array}{ll}
TR' \rightarrow DR\ TR & TR \rightarrow tnc \\
DR' \rightarrow SR\ DR & DR \rightarrow dom \\
TR'' \rightarrow TR'\ TR & SR \rightarrow sub \\
TR'' \rightarrow TR\ TR' &
\end{array}
$$

We show an example of syntactic tree by our rules in Fig. 1.

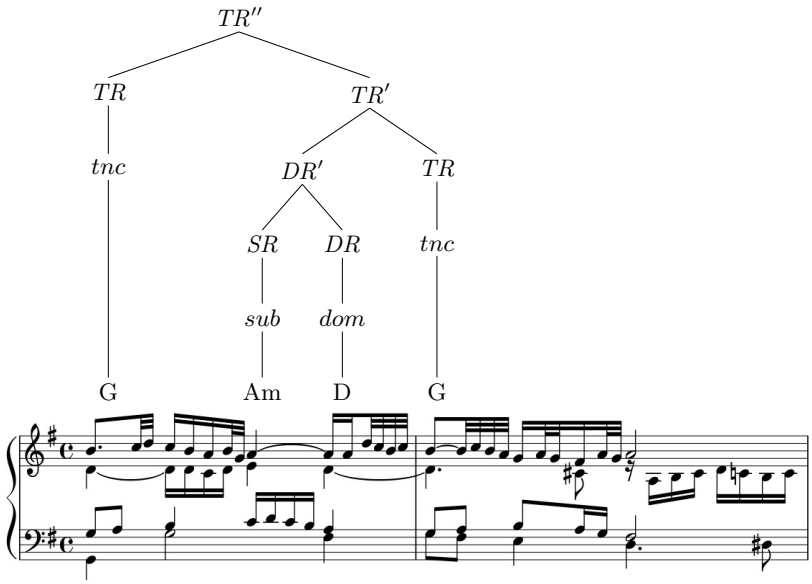


Fig. 1. J. S. Bach: "*Liebster Jesu, Wir Sind Hier*", BWV731

3 Semantics in Logic

3.1 Modal Logic

A modal logic consists of such syntax as follows.

$$\phi := p \mid \neg\phi \mid \phi \vee \phi \mid \Box\phi$$

where $\Box\phi$ represents, intuitively, 'necessarily ϕ.' In order to give a formal semantics to modal logic, we provide a *Kripke* frame which is a triplet $\mathcal{M} = \langle W, R, V \rangle$. W is a set of possible worlds and R is an accessibility relation among worlds, and

when w' is accessible from w we write wRw'. V is a valuation for each atomic proposition. Employing the negation connective, we also introduce[5]

$$\phi \wedge \phi \equiv \neg(\neg\phi \vee \neg\phi),\ \phi \supset \psi \equiv \neg\phi \vee \psi,\ \text{and}\ \Diamond\phi \equiv \neg\Box\neg\phi.$$

Now, we give the semantics of '$\Box$' operator, and its *dual* operator '$\Diamond$', as follows.

$$\mathcal{M}, w \models \Box\phi \textit{ iff } \text{ for all } w'(wRw') \in W,\ \mathcal{M}, w' \models \phi$$
$$\mathcal{M}, w \models \Diamond\phi \textit{ iff } \text{ there exists } w'(wRw') \in W \text{ and } \mathcal{M}, w' \models \phi$$

that is, $\Box\phi$ holds in w if and only if ϕ holds in all the accessible worlds w' from w, and $\Diamond\phi$ holds in w if and only if there exists an accessible world w' where ϕ holds.

We also read '$\Box\phi$' as '(a certain agent) knows ϕ' in epistemic logic since the knowledge of agent is a part of propositions which are true to her after considering (accessing) every possibility. In deontic logic '$\Box\phi$' means 'ϕ is obligatory.' In our study, this formula means ϕ persists for a certain extent of time.

Topological Semantics In Kripke semantics, we have presupposed that a set of possible worlds is a finite number of discrete ones, however, we can naturally extend the notion to a continuous space. Given a mother set X, let $\mathcal{P}(X)$ be the powerset of X and let $\mathcal{T} \subset \mathcal{P}(X)$. We call $\mathcal{T}$ is a topology if it satisfies

1. for $\mathcal{O}_1, \mathcal{O}_2 \in \mathcal{T}$, $\mathcal{O}_1 \cap \mathcal{O}_2 \in \mathcal{T}$.
2. for any (possibly infinite) index set Λ, if each $\mathcal{O}_i (i \in \Lambda) \in \mathcal{T}$, $\cup_{i\in\Lambda}\mathcal{O}_i \in \mathcal{T}$.
3. $X, \emptyset \in \mathcal{T}$.

A topological semantics of modal logic adopts a topology $\mathcal{T}$ instead of accessibility relation R, rereading possible worlds as geometric *points* in a more general way. A member of $\mathcal{T}$ is an open set that does not have intuitively hard boundaries in a metric space, or an *open ball*.

$$\mathcal{M}, w \models \Box\phi \textit{ iff } \text{there exists } \mathcal{O} \in \mathcal{T}(w \in \mathcal{O}),\ \text{for all } w' \in \mathcal{O}, \mathcal{M}, w' \models \phi \quad (1)$$
$$\mathcal{M}, w \models \Diamond\phi \textit{ iff } \text{for all } \mathcal{O} \in \mathcal{T}(w \in \mathcal{O}),\ \text{there exists } w' \in \mathcal{O}, \mathcal{M}, w' \models \phi \quad (2)$$

that is, $\Box\phi$ holds if and only if there is a set in $\mathcal{T}$ in which at all the points ϕ holds, while $\Diamond\phi$ holds if and only if in any set in $\mathcal{T}$ there is a point at which ϕ holds.

Among various arbitrary topologies, the *discrete topology* consists of all the subset of mother set X, that is, $\mathcal{T} = \mathcal{P}(X)$. Then, every point is distinguished from each other.

When $w \models \phi$, in any $\mathcal{O}$ including w, $w \models \phi$, so that Axiom

$$\phi \supset \Diamond\phi \qquad (\mathbf{T}^*)$$

inevitably holds.[6]

[5] Note that '$\supset$' is not a set inclusion but a logical implication.
[6] Axiom $\mathbf{T}$ is $\Box\phi \supset \phi$ and $\mathbf{T}^*$ is its dual form.

Neighborhood Semantics We call $N(w)$ is a set of neighborhoods of w. When we are given neighborhoods as a set of open balls for each point, we can construct a topology by a technique called *filtration*.[7] The semantics of modal operators are defined as follows.

$$\mathcal{M}, w \models \Box\phi \textit{ iff } [\![\phi]\!] \in N(w)$$
$$\mathcal{M}, w \models \Diamond\phi \textit{ iff } [\![\phi]\!]^c \notin N(w)$$

where $[\![\phi]\!] = \{w \mid \mathcal{M}, w \models \phi\}$, and $[\![\phi]\!]^c = W \setminus [\![\phi]\!]$ is the complement set of W.

3.2 Temporal Logic

Now we translate modal logic to temporal logic.

- Let a point (a world in Kripke semantics) be a time point.
- Let an open ball be an open time interval.

Here, we provide two different relations among intervals.

$$\tau_1 \preceq \tau_2 \colon \tau_1 \text{ temporally precedes } \tau_2$$
$$\tau_1 \subseteq \tau_2 \colon \tau_1 \text{ is temporally included in } \tau_2$$

However, this formalism is confusing since our intuitive interpretation of the precedence relation is for time points while the inclusion relation refers to intervals. In order to treat points and intervals impartially, we employ topological semantics, as follows.

Let '$\prec$' be the total order of τ_i's. Then, the *convex* open ball is such that for $\tau, \tau' \in \mathcal{O}$ any τ'' $(\tau \prec \tau'' \prec \tau) \in \mathcal{O}$. Then, let $\mathcal{T}_\prec$ be

$$\mathcal{T}_\prec = \{\mathcal{O} \in \mathcal{T} \mid \mathcal{O} \text{ is convex}\}.$$

From now on, we regard a convex open ball is a time interval. In short, we write (τ_i, τ_j) to represent the minimum convex open ball including τ_i and τ_j; if an open ball consists of a single pitch event τ_i we write (τ_i). Meanwhile, $\{\tau_i, \tau_j\}$ represents a set consists of two time points and we write $\{\tau_1, \tau_2, \cdots\} \models \phi$ for $\tau_i \models \phi$ $(i = 1, 2, \cdots)$.

4 Logical Annotation in Music

Our task in this paper is to annotate the syntactic tree by such formulae and to identify the harmonic regions among the time intervals. To be precise, according to a pitch event found at each time point τ, we aim at fixing the neighborhood $N(\tau)$ to validate the formula at τ, and name those intervals as regions.

[7] $N(w)$ is a *filter* when $N(w)$ is a set of neighborhoods of w, and for any $U \in N(w)$ there exists $V(\subset U) \in N(w)$ and for all $w' \in V, U \in N(w')$.

4.1 Logical Formula for Syntactic Rule

Now we give the syntax of logical formulae as follows.

$$\phi ::= f(x) \mid f(c) \mid \neg\phi \mid \phi \vee \phi \mid \forall x\phi \mid \Box\phi$$

where $f \in \mathbb{F}, c \in \mathbb{O}$ supplemented with chord name variables. We provide other logical connectives $\wedge$ and $\supset$ as before, as well as modal operator $\Diamond$ and $\exists x\phi \equiv \neg\forall x\neg\phi$. We often write $\tau \models \phi$ when ϕ holds at time τ omitting the frame $\mathcal{M}$ for simplicity; notice that ϕ is not a pitch event itself but a proposition with function name as predicate.

In order to implement a progression model instead of a generation model, we employ Earley's algorithm [7]. Let us consider a generative (production) rule $A \to B\ C$. Then the rule is evoked and executed in the following process.

1. Observe an pitch event with a harmonic function.
2. Find such generative rule(s) that the first item of the right-hand side B matches the observation.
3. To complete the upper category A residing at the left-hand side of '$\to$', expect the second item in the right-hand side C of the rule.

We provide logical formulae with variables, corresponding to each syntactic rule, including

$\Diamond f(x) \quad \cdots$ There is a pitch event x with function f in the neighborhood.
$\Box\Diamond f(x) \cdots$ Anywhere in some region ('$\Box$'), we can access ('$\Diamond$') $f(x)$.

Tentatively instead of $\mathbb{F} = \{tnc, dom, sub\}$, let $head, spc, cmp \in \mathbb{F}$, standing for X, *Spec*, and *Comp*, respectively, in X-bar rules. The syntactic head must *exist* while specifiers and complements may arbitrarily be accompanied. Then, to annotate each X-bar rule with logical formula, we propose the following quantifications.

$$\begin{cases} \exists x[head(x) \wedge \cdots] \\ \forall x[spc(x) \supset \cdots] \\ \forall x[cmp(x) \supset \cdots] \end{cases}$$

Here, we combine the above components, considering *spc* and *cmp* appear before and after the head, respectively, in accordance with the usual *dual* relation of $\forall x[\phi \supset \psi]$ versus $\exists x[\phi \wedge \psi]$,[8] as follows.

– For $X' \to Spec\ X$ with $X \to head$ and $Spec \to spc$,

$$\forall x[spc(x) \supset \exists y[\Diamond head(y) \wedge \Box\Diamond head(y)]]. \tag{3}$$

– For $X' \to X\ Comp$ with $X \to head$ and $Comp \to cmp$,

$$\exists x[head(x) \wedge \forall y[\Diamond cmp(y) \supset \Box\Diamond head(x)]]. \tag{4}$$

[8] If we negate the whole $\forall x[\phi \supset \psi]$ as $\neg\forall x[\phi \supset \psi]$ we obtain $\exists x[\phi \wedge \neg\psi]$, and vice versa.

The intuitive reading is that if we observe $spc(x)$ there exists a head event $head(y)$ in the neighborhood as $\Diamond head(y)$. The combined events could form a region (a temporal extent), that is a neighborhood $\mathcal{O}$ and $\tau(\in \mathcal{O}) \models \Diamond head(y)$, that is $\tau \models \Box\Diamond head(y)$. If we observe $head(x)$ first it is possibly accompanied by $cmp(y)$ in the neighborhood, that is $\Diamond cmp(y)$. Then, there should be a wider region in which $\Box\Diamond head(x)$ holds.

Example 1 For $DR' \rightarrow SR\ DR$, since d is the head, we expect that the following formula according to (3) would be satisfied.

$$\forall x[sub(x) \supset \exists y[\Diamond dom(y) \wedge \Box\Diamond dom(y)]].$$

Let C be a subdominant in G_{maj} found at τ_2 as $\tau_2 \models sub(C)$. Though $\tau_2 \models \Diamond dom(y)$ y is not bound yet. Next, we envisage we find *dominant* in the neighborhood of τ_2. Let $\tau_3 \models dom(D)$. This meets the expectation at τ_2 as $\tau_2 \models \Diamond dom(D)$. Also by T*, $\tau_3 \models \Diamond dom(D)$ and thus $\{\tau_2, \tau_3\} \models \Box\Diamond dom(D)$. To be reminiscent of original region names, (τ_2) is SR and (τ_2, τ_3) is DR, respectively. ■

Example 2 For $TR'' \rightarrow TR'\ TR$, we expect

$$\exists x[tnc(x) \wedge \forall y[\Diamond tnc(y) \supset \Box\Diamond tnc(x)]]$$

would be satisfied according to (4). Let $\tau_7 \models tnc(G)$. If $\tau_9 \models tnc(E_m)$ we can consider E_m is a complement of head G, and $\{\tau_7, \tau_8, \tau_9\} \models \Box\Diamond tnc(G)$. Then, (τ_7, τ_9) becomes TR. ■

We summarize our logical formulae corresponding to abridged X-bar syntactic rules in Table 1.

$TR' \rightarrow DR\ TR$	$\forall x[dom(x) \supset \exists y[\Diamond tnc(y) \wedge \Box\Diamond tnc(y)]]$
$DR' \rightarrow SR\ DR$	$\forall x[sub(x) \supset \exists y[\Diamond dom(y) \wedge \Box\Diamond dom(y)]]$
$TR'' \rightarrow TR'\ TR$	$\exists x[tnc(x) \wedge \forall y[\Diamond tnc(y) \supset \Box\Diamond tnc(x)]]$
$TR'' \rightarrow TR\ TR'$	$\forall x[tnc(x) \supset \exists y[\Diamond tnc(y) \wedge \Box\Diamond tnc(y)]]$
$TR \rightarrow tnc$	$\exists x[tnc(x) \wedge \Box\Diamond tnc(x)]$
$DR \rightarrow dom$	$\exists x[dom(x) \wedge \Box\Diamond dom(x)]$
$SR \rightarrow sub$	$\exists x[sub(x) \wedge \Box\Diamond sub(x)]$

Table 1. Logical formulae for X-barred GSM

4.2 Valuation with Key

In general, a logical formula ϕ in modal logic is evaluated, that is to decide true or false, given a Kripke frame or simply *model* $\mathcal{M}$ with possible world w as follows.

$$\mathcal{M}, w \models \phi.$$

In our case, a possible world is a time point τ and the frame $\mathcal{M}$ is either topology $\mathcal{T}$ or the neighborhoods, thus

$$\mathcal{M}, \tau \models \phi.$$

Especially, when ϕ includes modal operators, our task has been to fix $N(\tau)$ so as to validate ϕ.

Besides, we need key information in our context of music on the left-hand side of '$\models$'. When we observe a set of notes {D, E♯, A} at a certain time point in a music piece, we recognize the pitch event as chord D, though we may not yet be aware of the scale degree and its key. A chord in $\mathbb{O}$ possesses a function in $\mathbb{F}$ dependent on a key. If the context is G major, we can assign the *dominant* function to this D. That is, $tnc(\mathrm{G}), dom(\mathrm{D}), sub(\mathrm{C})$ all hold in G major while not in C major.

Since we have provided a topology or a neighborhood, we need to detail the left-hand side of '$\models$', as

$$\mathrm{Gmaj}, \mathcal{M}, \tau \models dom(\mathrm{D}).$$

Or, in general, $K.\mathcal{M}, \tau \models \phi$ where $K \in \mathbb{K}$. For simplicity, in the remaining discussion we omit such key information as long as there is no confusion.

4.3 Detailed Analysis

As an example, we detail the analysis for the beginning part of "*Liebster Jesu, Wir Sind Hier*" of Fig. 1, in Table 2.

Fig. 2 shows the bottom half of the processes in Table 2, (i) the recognition of subdominant C evokes the expectation of dominant, (ii) the expectation to detect a dominant, (iii) The detection of dominant D satisfies the *DR* and evokes another expectation of tonic, (iv) the expectation towards the tonic, and (v) The final G satisfies the tonic function and the completion of the tree. In Fig. 3, we show the suggested regions by our analysis.

5 Discussion

In this paper, we have proposed the logical semantics of music. We employed the X-bar theory and reinterpreted the existing grammar rules according to the notation. Thereafter, we have shown the principles of producing logical formulae. Finally, we have shown the analysis example based on Generative Syntax Model.

Our contributions are two-fold. First, we have given a clear accessibility relations, from each pitch event to regions in the tree structure, in terms of logical

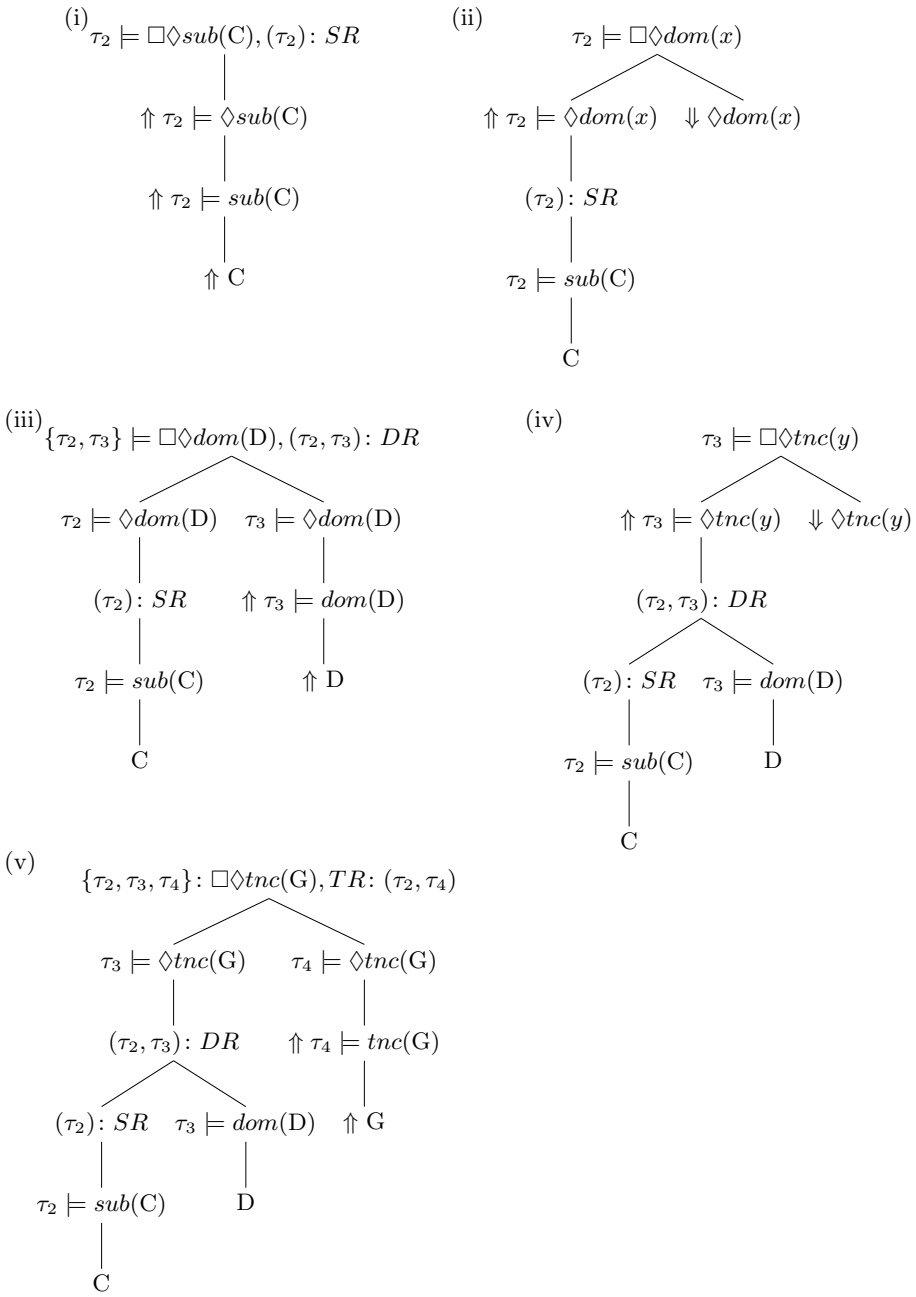


Fig. 2. Expectation-based Tree Construction of the bottom half of Table 2. The upward arrows mean the observation of pitch events in the bottom-up process and the downward arrows show the expectation. SR, DR, TR represent regions.

$\tau_1 \models$	$tnc(\mathrm{G}_1)$	We observe chord G_1 at τ_1 as tonic.
	$\Diamond tnc(\mathrm{G}_1)$	By $\mathbf{T}^*$.
	$\Diamond tnc(x)$	Still another tonic events may follow.
	$\Box\Diamond tnc(\mathrm{G}_1)$	G_1 is the head.
$\tau_2 \models$	$sub(\mathrm{C})$	We observe C at τ_2 as subdominant. (i)
		as a result, $\tau_1 \not\models \Diamond tnc(x)$ and (τ_1) becomes TR.
	$\Diamond sub(\mathrm{C})$	By $\mathbf{T}^*$.
	$\Box\Diamond sub(\mathrm{C})$	(τ_2) becomes SR.
	$\Diamond dom(y)$	We expect a dominant coming. (ii)
$\tau_3 \models$	$dom(\mathrm{D})$	We observe D as dominant at τ_3. (iii)
	$\Diamond dom(\mathrm{D})$	By $\mathbf{T}^*$.
		$\tau_2 \models \Diamond dom(\mathrm{D})$ by $y = \mathrm{D}$.
	$\Diamond tnc(z)$	Also, we expect a tonic follows. (iv)
	$\Box\Diamond dom(\mathrm{D})$	D is the head.
		(τ_2, τ_3) becomes DR.
$\tau_4 \models$	$tnc(\mathrm{G}_2)$	We observe G_2 as tonic. (v)
	$\Diamond tnc(\mathrm{G}_2)$	By $\mathbf{T}^*$.
		$\tau_3 \models \Diamond tnc(\mathrm{G}_2)$ by $z = \mathrm{G}_2$.
	$\Box\Diamond tnc(\mathrm{G}_2)$	G_2 is the head.
		(τ_2, τ_4) becomes TR.
		Combining (τ_1) TR, (τ_1, τ_4) becomes TR.

Table 2. Chronological recognition of pitch events; two G's at the beginning and at the end are distinguished by indices. (i)–(v) correspond to the subtrees in Fig. 2. We have omitted the key as well as $\mathcal{M}$ on the left-hand side of '$\models$' for simplicity.

formulae. Therefore, we can annotate each node (branching point) in the tree with formulae and can clarify what has been expected at that time. As a result, we have distinguished the time points and regions in a rigorous way.

Second, the top-down procedure of generation, consisting of a chaining of multiple generative rules can be reinterpreted to the progressive model. When we listen to music, or compose a music, we construct the music score in our mind from in the sequential way. By our reinterpretation, now each rule works as an expectation-realization model.

Finally, we discuss various future tasks. (a) We have divided a music piece in a disjoint, hierarchical regions in accordance with GTTM's grouping analysis [10], however, a short passage, *e.g.*, pivot chords, may be interpreted multiple ways. We need to equip our theory with flexible overlapping regions. (b) Though we have considered the semantics by topology, this might be unnecessarily complicated. We may simply need a future branching model with the necessary expectation **G** and the possible expectation **F** in Computational Tree Logic (CTL) [5]. (c) We need also to consider how such expectation corresponds to the existing implication–realization model (I-R model) [15], or to tension–relaxation structure in GTTM [10] in general.

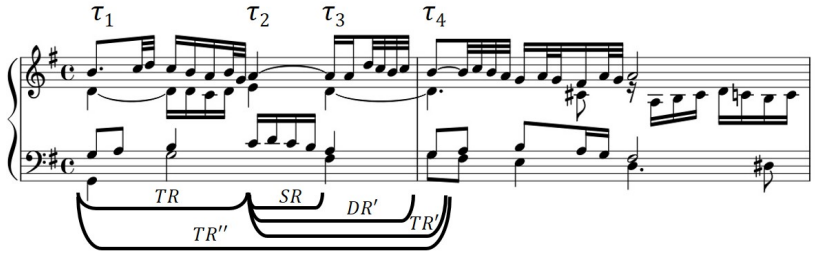


Fig. 3. Time points and regions in the beginning of BWV731

References

1. Chomsky, N.: *Sytactic Structures*, Mouton & Co. (1957)
2. Chomsky, N.: *Aspects of the Theory of Syntax*, The MIT Press (1965)
3. Chomsky, N.: Remarks on nominalization. In: Jacobs, R. and Rosenbaum, P. (eds.): *Reading in English Transformational Grammar*, 184-221 (1970).
4. Chomsky, N.: The Minimalist Program, The MIT Press (1995)
5. Clarke, E. M, Emerson, E. A.: Design and synthesis of synchronisation skeletons using branching time temporal logic, Logic of Programs, Proceedings of Workshop, Lecture Notes in Computer Science, Vol. 131. pp52–71 (1981)
6. Dowty, D. R., Wall, R. E., Peters, S.: *Introduction to Montague Semantics*, D. Reidel Publishing Company (1981)
7. Earley, J.: An efficient context-free parsing algorithm, *Communications of the Association for Computing Machinery*, 13:2:94-102 (1970)
8. Granroth-Wilding, M., Steedman, M.: A robust parser-interpreter for jazz chord sequences. Journal of New Music Research **43**, 354–374 (2014)
9. Gollin, E., Rehding, A. eds.: *The Oxford Handbook of Neo-Riemannian Music Theories*, Oxford (2011)
10. Lehrdahl, F., Jackendoff, R.: *A Generative Theory of Tonal Music.* The MIT Press (1983)
11. Hamanaka, M., Tojo, S., Hirata, K.: Implementing a general theory of tonal music. *Journal of New Music Research* **35**(4), 249–277 (2007)
12. Koelsch, S.: *Brain and Music*, John Wiley & Sons, Ltd. (2015)
13. Meyer, L.E.: Meaning in music and information theory. *The Journal of Aestheticsex and Art Criticism* **15**(4), 412–424 (1957)
14. Narmour, E.: *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*, The University of Chicago Press (1990)
15. Narmour, E.: *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*, The University of Chicago Press (1992)
16. Rohmeier, M.: Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music* **5**(1), 35–53 (2011)
17. Tojo, S., Oka, Y., Nishida, M.: Analysis of chord progression by hpsg. In: Proceedings of 24th IASTED international conference on Artificial Intelligence and its applications (2006)
18. Wallin, N. L., Merker, L. and Brown, S. (eds.).: *The Origins of Music.* The MIT Press (2000)
19. Winograd, T.: Linguistics and the computer analysis of tonal harmony, *Journal of Music Theory*, **12**(1) (1968)

John Cage's Number Pieces, a geometric interpretation of "time brackets" notation

Benny Sluchin[1] and Mikhail Malt[2],

[1] IRCAM/EIC
[2] IRCAM/IReMus
sluchin@ircam.fr, mikhail.malt@ircam.fr

Abstract. Conceptual musical works that lead to a multitude of realizations are of special interest. One can't talk about a performance without considering the rules that lead to the existence of that version. After dealing with similar works of open form by Iannis Xenakis, Pierre Boulez and Karlheinz Stockhausen, the interest in John Cage's music is evident. His works are "so free" that one can play any part of the material; even a void set is welcomed. The freedom is maximal and still there are decisions to consider in order to make the piece playable. Our research was initially intended to develop a set of conceptual and software tools that generates a representation of the work as an assistance to performance. We deal here with the *Number Pieces* Cage composed in the last years of his life. Over time, we realized that the shape used to represent time brackets, brought important information for the interpretation and musical analysis. In the present text, we propose a general geometric study of these time brackets representations, while trying to make the link with their musical properties to improve the performance.

Keywords: Computer Aided Performance, Notation, Musical Graphic Representation

1 Introduction

The interpreter who approaches the music of John Cage composed after the middle of the 20th century is often disconcerted by a great freedom of execution, associated with a set of precise instructions. The result is that, each time, the musician is led to determine "a version," and to decide on a choice among the free elements proposed by the piece. A fixed score is thus created, which can be used several times. The musician interprets "his version" while thinking that it conforms to the composer's intentions. But in fact, most works of Cage composed after the 1950s should not be preconceived, prepared, "pre-generated" for several executions. Each interpretation should be unique and "undetermined." It is in this sense that the use of the computer can help the performer: a program will allow the latter to discover without being able to anticipate what and when he plays. The performance of the work thus escapes the intention of the musician to organize the musical text.

2 John Cage's Number Pieces

The corpus of John Cage's late compositions (composed between 1987 and 1992) is known today as *Number Pieces*. Each work is named after the number of musicians involved; and the exponent indicates the order of the piece among the other compositions containing the same number of musicians [1].

Silence and Indeterminacy

In the course of his creative research as a composer, Cage has laid down essential structural elements. Thus, silence has been posited as an element of structure to be thought of in a new and positive way; not as an absence of sound, but as a diachronic element, a presence, an acoustic space. This innovative work concerning silence has itself evolved: at first it was conceived as giving the work its cohesion by alternating with sound, then Cage extended the reflection to a spatial conception: the silence is composed of all the ambient sounds which, together, form a musical structure. Finally, silence was understood as "unintentional," sound and silence being two modes of nature's being unintentional [2].

Moreover, in this desire to give existence to music by itself, Cage has resorted to various techniques of chance in the act of composition and principles of performance.

The principles of indetermination and unintentionality go in that direction. The principle of indetermination leads the musician to work independently from the others, thus introducing something unexpected in what the musical ensemble achieves. The performer, unaware of the production of his fellow musicians, concentrates on his own part and on the set of instructions. This requires great attention, even if the degree of freedom of the playing is high [3].

Time Brackets

In Cage's Number Pieces each individual part contains musical events with *time brackets*. Generally, an event consists of a score endowed with two pairs of numbers: time brackets (Fig. 1).

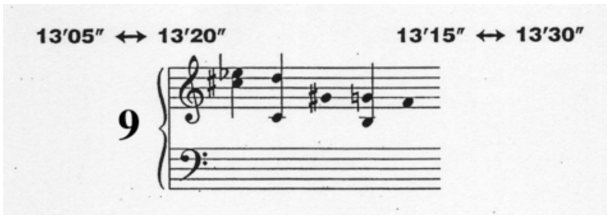


Fig. 1. John Cage's *Two*[5], piano, 9th event

This gives the interpreter lower and upper-time bounds to begin and end each event. The composition has a defined total duration and the events are placed inside a pair of

the *time brackets*. Although there are only individual parts, a score for the group is implicitly present and leads to a form.

Earlier research

In previous work [8] we modeled these time brackets by parallelograms (see Figures 2 and 3) to build computer interfaces for interpretation assistance in the context of Cage's *Two*[5] (Fig. 2).

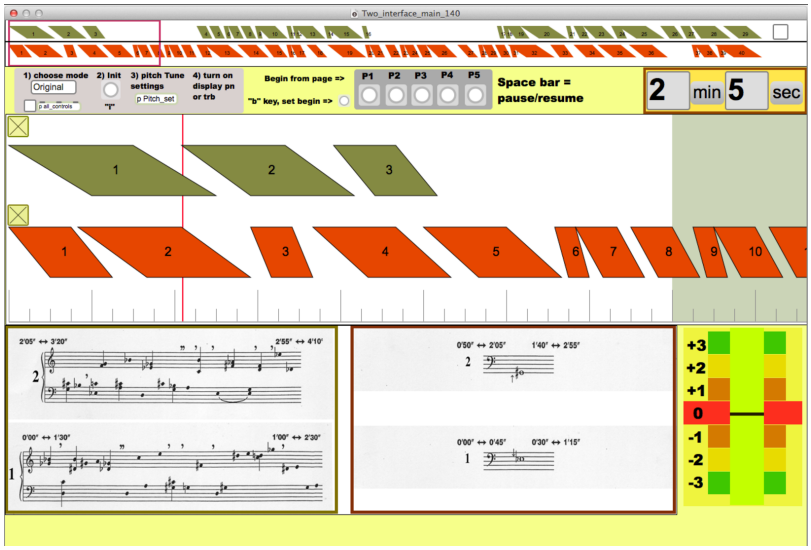


Fig. 2. Cage's *Two*[5] main computer interface

Over time ([9], [10], [11]), we realized that the shape used to represent time brackets, brought important information for the interpretation and musical analysis. The unusually long duration of this piece, 40 minutes, and the use of time brackets show that the temporal question, and its representation, is essential in the Number Pieces, in general, and in *Two*[5] in particular.

The computer interface whose use has become obvious, has created for us a climate of confidence in our relationship to the piece. Random encounters of synchronicity as well as intervals bring unexpected situations…[12]

In the present text, we propose a general geometric study of these time brackets representations, while trying to make the link with their musical properties to improve the performance.

3 The Geometry of Time Bracket

The first step in the process was to model a graphic representation of each part as a succession of musical events in time. For this purpose, the temporal structure of the piece has been represented as quadruples on a timeline. $(s_l(k),\ s_u(k),\ e_l(k), e_u(k))$.

In order to place an event *k* on the timeline, time brackets are defined as quadruples to indicate the time span allocated to it. Each quadruple consists of two pairs. More precisely, each pair gives the interpreter lower and upper time bounds to start $(s_l(k),\ s_u(k))$ and to end $(e_l(k),\ e_u(k))$. Theses closed time intervals give to the performer, a choice of the pair $(s(k),\ e(k))$, where $(s_l(k) \leq s(k) \leq s_u(k))$ and $(e_l(k) \leq e(k) \leq e_u(k))$. One could choose the starting time $(s(k))$, while performing and, then accordingly, the end time $(e(k))$. This is the way one would employ when actually performing the work.

To obtain a graphic representation of each event in time we consider the quadruple:

$$(s_l(k),\ s_u(k),\ e_l(k), e_u(k))$$

where $(s_l(k),\ s_u(k))$ is the *Starting Time Zone* and $(e_l(k),\ e_u(k))$ the *Ending Time Zone*. As the two intervals have, in our case, a designed superposition, we prefer to distinguish starting and ending zones by using two parallel lines (Fig. 3).

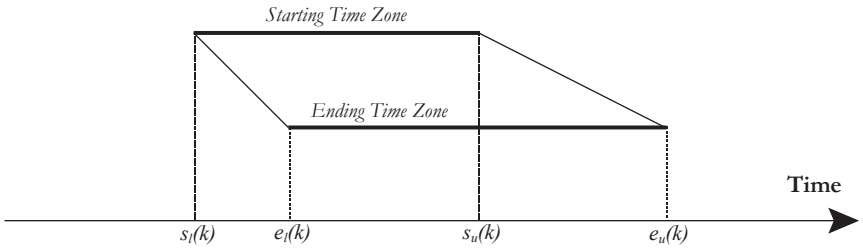


Fig. 3. Graphic representation for a generic time event

The graphic event obtained by connecting the four points has a quadrilateral shape. The height has no particular meaning. The *starting duration* $\delta_s(k)$ is defined as the difference: $(s_u(k) - s_l(k))$, which is the time span the performer has to start the event. In the same way the *ending duration* $\delta_e(k)$ will be the time span given to end the event $(e_u(k) - e_l(k))$. In the general case, these values are not the same, and the form we get is asymmetrical. When dealing with Cage's *Number Pieces*, one generally has: $\delta_s(k) = \delta_e(k)$, both durations are the same, and the figure to represent an event is a trapezoid (Fig. 4). This is the case in the majority of the corpus we are treating. Special cases will be mentioned later on.

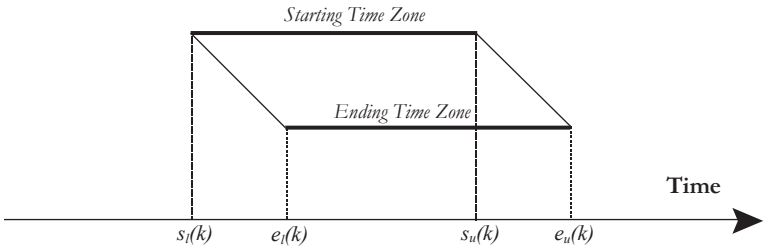


Fig. 4. Graphic representation for a time event in Cage's *Number Pieces*

There is mostly an overlapping of the two time zones, $(s_l(k),\ s_u(k))$ and $(e_l(k),\ e_u(k))$ but it can happen that those are disjoined. We can define a variable $\gamma(k)$ where: $s_l(k) + \gamma(k) = e_l(k)$. In Cage's *Number Pieces*, $\gamma(k)$ depends generally on the event duration. Thus, we don't have a big variety of forms. For example, in *Five*[3], we have only 4 different time brackets sorts, for a total number of 131 events for the five instruments and $\gamma(k) = \frac{2}{3}\delta\ (k)$ for all quadruples.

We make a distinction between a *generic musical event* and a *real* (or determined) *musical event*. A real musical event is the one whose starting points (*s*) and end points (*e*) are defined, that is, where there is a concretization of choice. One could represent this by a straight line from $s(k)$ to $e(k)$ (Fig. 5).

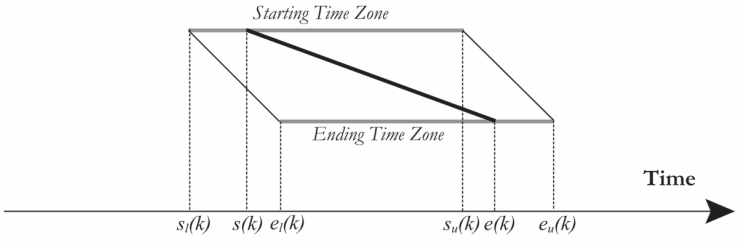


Fig. 5. A real music event represented by a straight line, joining the starting to ending time zones

There are certain properties of a generic event that can easily be deduced from the trapezoidal graphic representation:

1. The *starting* or *ending durations*: $\delta_s(k)$ or $\delta_e(k)$ are a kind of a nominal duration that Cage gives to an event.

2. The maximum duration, $e_u(k) - s_l(k) = \delta_{max}(k)$, is the maximum length (duration) an event can have.

3. The fact that, $s_u(k) > e_l(k)$ means that we can choose a starting point *s(k)* placed after the end, which leads to an empty musical event ∅ (an important idea of Cage: he often indicates that the artist can choose, all of, a part of, or nothing of the material placed at its disposal). In this case, *s(k) > e(k)*.

4. An alternative way to present a quadruple will be: $(s_l(k), \delta_s(k), \delta_e(k), \gamma(k))$ where $\gamma(k)$ is the value previously discussed. This representation can easily display the regularity in the time brackets construction (Fig. 6). It is easy to see that

$$\delta_{max}(k) = \frac{(\delta_s(k) + \delta_e(k))}{2} + \gamma(k).$$

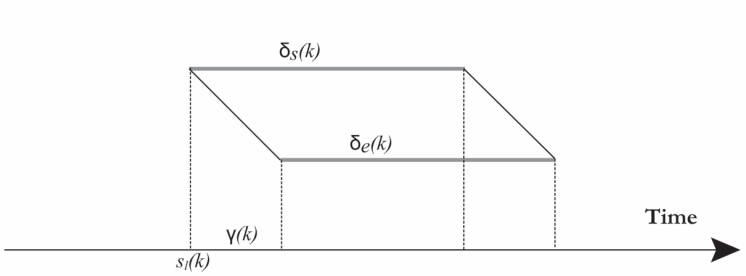


Fig. 6. An event represented as $(\boldsymbol{s_l(k)}, \boldsymbol{\delta_s(k)}, \boldsymbol{\delta_e(k)}, \boldsymbol{\gamma(k)})$

5. An implicit parameter that is important is the straight line's slope of the concrete event (Fig. 5). This value is inversely proportional to the concrete event duration. The slope is strongly related to performance: it shows how much time the performer has for a particular event *k*. In regard to a wind instrument part, often only composed by held notes, knowledge of this parameter allows the artist to better manage his air capacity, in order to respect the composer's indications. As far as the pianist is concerned, the slope gives some information that allows him to manage his interpretation with reference to the time indications. When the straight line of a concrete event is close to the vertical, the event will be short and concentrated.

The relationships of the generic events

Concerning the placement of two contiguous events *k* and *k+1* we can define a variable $\varepsilon(k)$, the gap between the elements *k* and *k+1* where:
$\varepsilon(k) = s_l(k+1) - e_u(k)$ (Fig. 7).

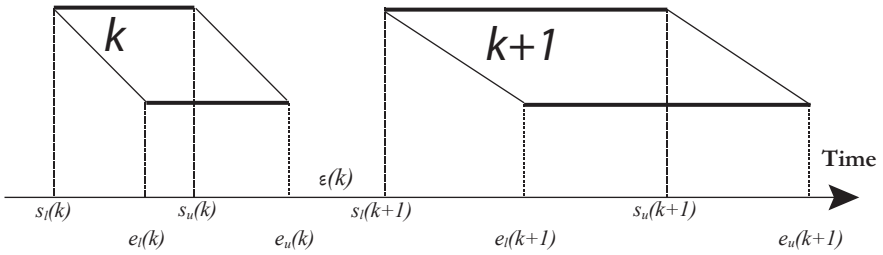


Fig. 7. $\boldsymbol{\varepsilon(k)}$, The gap between the elements *k and k+1*

We will observe five typical placements of two contiguous events.

1. $\varepsilon > 0$.

The two events are separated on the timeline. There is a minimum length of silence between the two events, which will probably be longer according to the choice of $e(k)$ and $s(k+1)$. In *Five*[3] for example, we have events 1 and 2 of violin 2 separated by more than 8 minutes, or 3 minutes between events 6 and 7 of violin 1. Here the piece could also be considered from the point of view of the relative density of the musical elements. One should mention the global statistical approach done elsewhere [4] [5].

2. $\varepsilon = 0$.

The two events are adjacent (Fig. 8).

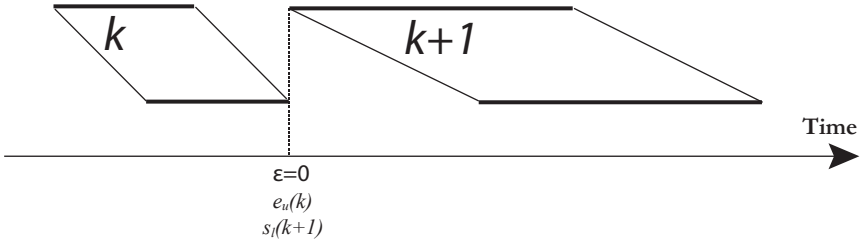


Fig. 8. $\varepsilon = 0$

Again, a gap may occur between the two events as the actual ending of event *k:* $e(k)$, and/or the actual starting of event *k+1,* $s(k+1)$ will differ from $e_u(k)$, and $s_l(k+1)$ correspondingly. For example, *Two*[5], trombone, events 21 and 22 (Fig. 9), events 27 and 28.

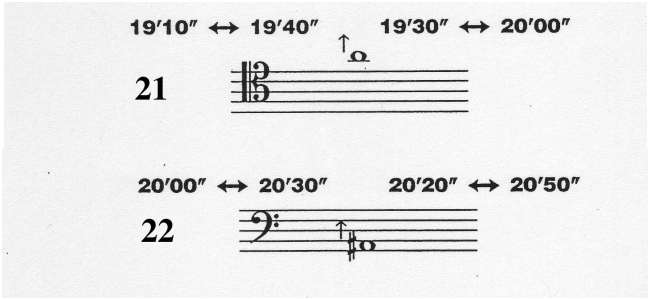


Fig. 9. *Two*[5], trombone, events 21 and 22

3. $\varepsilon < 0$.

In this case, the performer's opinion and attitude can determine the performance. There are many remarkable cases of interest in this situation; we could mention some cases that presently occur in Cage's *Number Pieces* (Fig. 10). For example, *Two*[5], trombone events 28 and 29, and piano events 6 and 7.

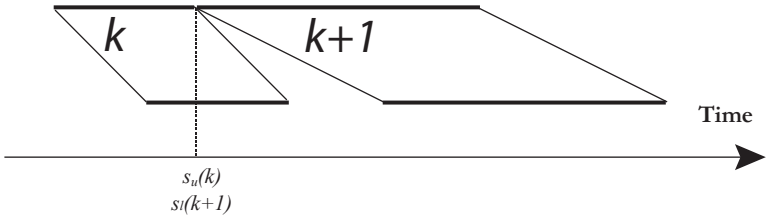


Fig. 10. $\varepsilon < 0$, $\boldsymbol{s_l(k+1) = s_u(k)}$

While performing event *k*, the player could start the event *k+1* when not yet ending event *k*. We can encounter a superposition as shown in Fig. 11. For example, *Two*[5], trombone events 37 and 38; piano events 9 and 10, events 12 and 13.

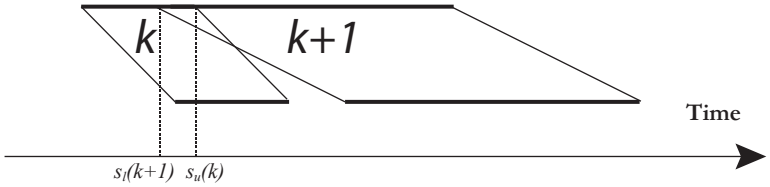


Fig. 11. $\varepsilon < 0$, $\boldsymbol{s_l(k+1) < s_u(k)}$

And even the same starting time for the two events: $s_l(k+1) = s_l(k)$ (Fig. 12). For example, *Two*[5], piano, events 14 and 15 (Fig. 13).

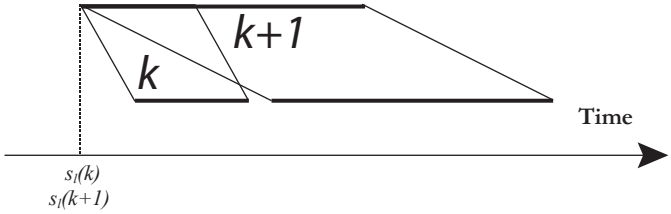


Fig. 12. $\varepsilon < 0$, $\boldsymbol{s_l(k+1) = s_l(k)}$

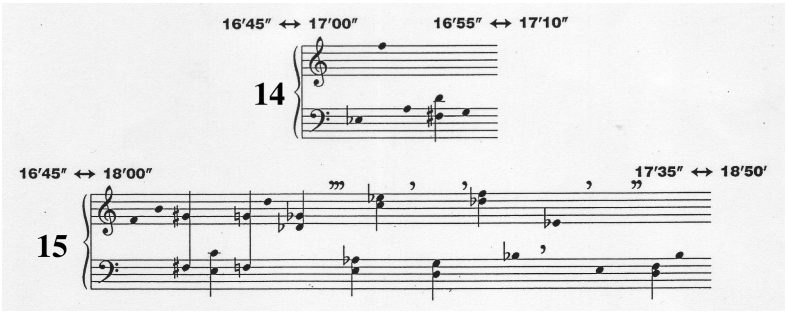


Fig. 13. *Two*[5], piano, events 14 and 15

As the events have an order given by Cage, one may assume that the sequence of events is to be respected. But the performer may consider mixing the two events and choosing the respective ending times, $e(k)$ and $e(k+1)$.

In some case one has the configuration shown in Fig. 14. For example, *Two*[5], trombone events 31 and 32, events 39 and 40.

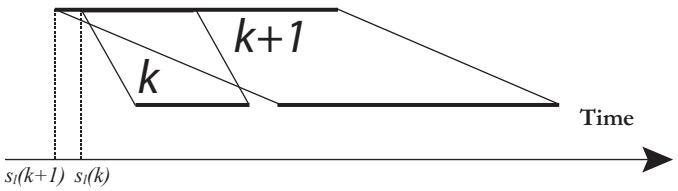


Fig. 14. $\varepsilon < 0$, $\boldsymbol{s_l(k+1) < s_l(k)}$

This may be a mistake, in calculation or in printing. Again, without change the order of events, one could start with the event *k*, and continue with the event *k+1*, mixing or separating. Starting with the event *k+1* would mean that mixing has to happen, or the event *k*, should be skipped, that an idea dear to Cage: the event *k* wouldn't be performed.

The presentation of the time brackets as geometric figures and the variables we have defined lead to calculate some constants related to each of the instruments involved. The *average filling rate* ($\overline{Fr}$) gives an indication of how much a particular instrument is present during the piece. This value will be the ratio of the sum of all the events' duration by the overall length of the work (Δ), where the event duration, $\delta(k)$, is the arithmetic mean between $\delta_s(k)$ and $\delta_e(k)$ (**1**).

$$\overline{Fr} = \frac{\sum_1^n \delta(k)}{\Delta} \qquad \textbf{(1)}$$

In the analog way, if we set: $\varepsilon(0)$ be the gap before the first event, and $\varepsilon(n)$ the gap after the last event *n*, the *average silence rate* ($\overline{Sr}$) will be the ratio of the sum of all the gaps between the events by the overall length of the work (**2**).

$$\overline{Sr} = \frac{\sum_0^n \varepsilon(k)}{\Delta} \quad (2)$$

These interesting values are based on the lengths of events, the gaps between them and their number, independent of the contents of the events.

If instead of using $\delta(k)$, the event duration, we consider $\delta_{max}(k)$, then:

$$\sum_{k=1}^{n} \delta_{max}(k) + \sum_{0}^{n} \varepsilon(k) \ = \Delta \quad (3)$$

4 Musical Analysis Application

Table 1 shows the values for the 21 events of violin 1 in *Five*[3], and the constants we just defined. The time values, onsets and durations, are defined in seconds.

Table 1. Data for *Five*[3], first violin

	s_l (k)	δ_s (k)	δ_e (k)	γ (k)	ε (k)		
1	10	45	45	30	10	$\Delta =$	**2400**
2	150	45	45	30	65		
3	215	30	30	20	-10	$\sum_{k=1}^{n} \delta(k) =$	**810**
4	290	45	45	30	25		
5	405	45	45	30	40		
6	465	45	45	30	-15	$\sum_{k=0}^{n} \varepsilon(k) =$	**1030**
7	740	45	45	30	200		
8	1225	45	45	30	410		
9	1315	15	15	10	15		
10	1325	45	45	30	-15	$\overline{Fr} =$	**0,3375**
11	1475	15	15	10	75		
12	1570	30	30	20	70	$\overline{Sr} =$	**0,4292**
13	1625	45	45	30	5		
14	1685	45	45	30	-15		
15	1865	30	30	20	105		
16	1900	45	45	30	-15		
17	2060	45	45	30	85		
18	2165	45	45	30	30		
19	2235	15	15	10	-5		
20	2245	45	45	30	-15		
21	2305	45	45	30	-15		

The following Table 2, compares these constants for the five instruments. We can observe how these two constants ($\overline{Fr}$ *and* $\overline{Sr}$) are strongly related to the presence of the instruments. For example, trombone will be more present, more active than the string instruments. One can see that $\overline{Sr}$ may be negative. This occurs when many of the events are superposed (All cases with $\varepsilon < 0$).

Table 2. Comparison values in *Five*[3]

	#Events	$\overline{Fr}$	$\overline{Sr}$
Violin 1	21	0.34	0.43
Violin 2	12	0.16	0.74
Viola	26	0.34	0.44
Violoncello	25	0.23	0.5
Trombone	47	0.74	-0.24

These values are clearly reflected in the form of the piece seen in the upper part of Fig. 15. We had implemented several models, some offline in "OpenMusic"[1] computer aided composition software, and in a real-time "Max" software [8]. Fig. 15 presents a generic computer interface we are exploring, to perform most part of Cage's *Number Pieces*.

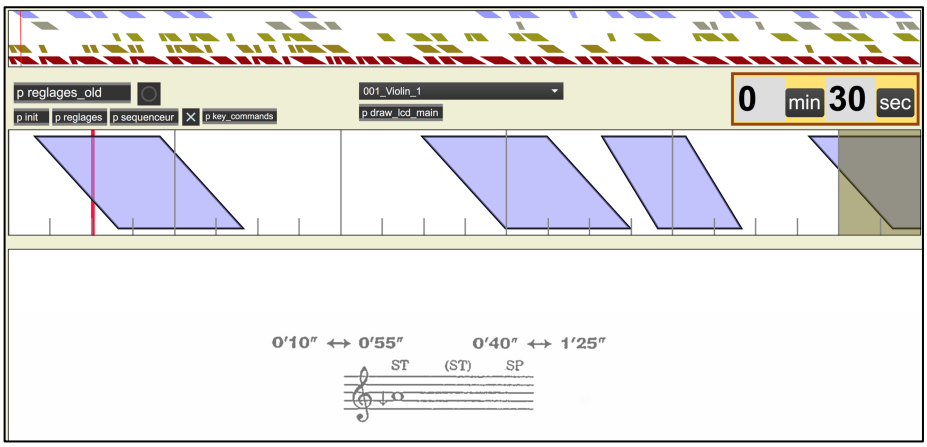


Fig. 15. Computer interface used for performing *Five*[3]

The medium part of this figure, displays one of the instruments chosen (here violin 1) and bottom part displays the musical score corresponding to the time (here 30 seconds after beginning). The global view displays a presentation of the entire duration of *Five*[3], using the trapezoidal event representation. It allows the performer to have a global view of the piece at a glance. As Cage mention about the context-specific character of his time-bracket notation:

> *Then, we can foresee the nature of what will happen in the performance, but we can't have the details of the experience until we do have it.* [6]

This global representation enables another perspective of the piece. The printed score orients a natural local view. More than being a graphic representation for each time bracket, it allows us to identify similarities between generic musical events. Fig. 16, a detail from Fig. 15, presents the first ten minutes of the global representation of *Five*[3].

[1] "OpenMusic" is a software developed by Ircam by Gerard Assayag, Carlos Augusto Agon and Jean Bresson. See: http://recherche.ircam.fr/equipes/repmus/OpenMusic/.

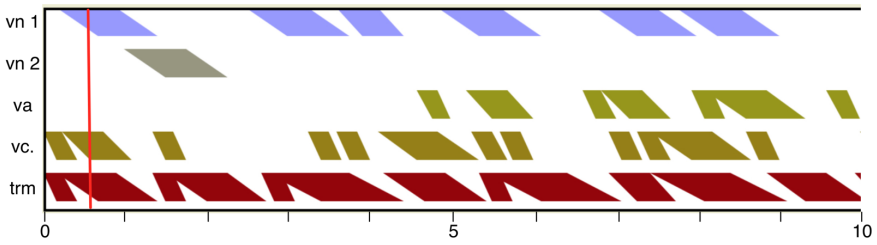


Fig. 16. The first ten minutes of the global representation in Cage's *Five*[3]

In an analog way Table 3 presents $\overline{Fr}$ and $\overline{Sr}$ constants for *Two*[5], and Fig. 17 shows the global structure of the piece. One can clearly distinguish the difference in the presence of the two instruments.

Table 3. Comparison values in *Two*[5]

	#Events	$\overline{Fr}$	$\overline{Sr}$
Piano	29	0.33	0.15
Trombone	40	0.46	-0.14

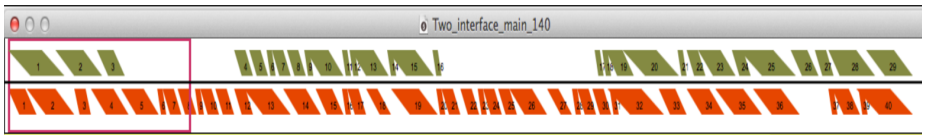


Fig. 17. Two[5] global structure

5 Conclusions

At the present time we work to offer the musicians a way to approach other pieces from the same family, constructing a generic interface. The task may be somewhat complicated. The works called *Number Pieces*, share the same principal described earlier, but often contain particularities and exceptions in the instructions for performance. The interface then has to be adapted to cover these.

The interface is a substitute to the printed score. It reveals the structure of the work and provides the performer with the tool to achieve the "meditative concentration" needed. The few instructions given by Cage are integrated in the interface.

Considering the graphic representation, we presented above, our main goal was to find geometric properties and strategies to enhance the performance of these pieces through computer interfaces. John Cage's works have been the target of our work for several years now. We have developed computer tools for the interface, and used it in practice. Both concerts and recordings have been the tests for the usefulness of the approach towards performance. The modeling process is transformed in a pragmatic analysis of the musical phenomena that leads us, step by step, to model some of Cage's

concepts. Mentioning first the *Concert for Piano and Orchestra* (1957), an earlier work that has become important step of his output [7]. Followed by two of his number pieces for a small number of performers [8]. These works were also the object of a recording and performance sessions ([9], [10], [11]).

References

1. Haskins, R.: The Number Pieces of John Cage. DMA dissertation, University of Rochester (2004). Published as Anarchic Societies of Sounds, VDM Verlag (2009).
2. Chilton, J. G.: Non-intentional performance practice in John Cage's solo for sliding trombone, DMA dissertation, University of British Columbia (2007).
3. Pritchett, J.: The Music of John Cage, Cambridge University Press, Cambridge (1993).
4. Popoff, A.: Indeterminate Music and Probability Spaces: the Case of John Cage's Number Pieces, MCM (Mathematics and Computing in Music) Proceedings, pp. 220--229. Springer Verlag (2011).
5. Popoff, A.: John Cage's Number Pieces: The Meta-Structure of Time-Brackets and the Notion of Time, Perspectives of New Music Volume 48, Number 1, pp. 65—83 (Winter 2010).
6. Retallack, J.: Musicage: Cage muses on words, art, music, Wesleyan university Press, p. 182 (1996).
7. Sluchin, B., Malt, M.: Interpretation and computer assistance in John Cage's Concert for piano and Orchestra (1957-58). 7th Sound and Music Conference (SMC 2010), Barcelona (21-24 July 2010).
8. Sluchin, B., Malt, M.: A computer aided interpretation interface for John Cage's number piece Two[5]. Actes des Journées d'Informatique Musicale (JIM 2012), pp. 211—218, Namur, Belgique (9–11 mai 2012).
9. Cage, J.: *Two*[5]. On John Cage Two[5] [CD]. Ut Performance (2013).
10. Cage, J.: *Music for Two*. On John Cage, Music for Two [CD]. Ut Performance (2014).
11. Cage, J.: *Ryoanji*. On John Cage, Ryoanji [CD]. Ut Performance (2017).
12. Rappaport, S., Sluchin, B.: On Panorama [CD]. Liner notes. Ut Performance (2019).

Modelling 4-dimensional Tonal Pitch Spaces with Hopf Fibration

Hanlin Hu and David Gerhard

Department of Computer Science, University of Regina, Regina SK, Canada
{hu263, gerhard}@cs.uregina.ca

Abstract. The question of how to arrange harmonically related pitches in space is a historical research topic of computational musicology. The primitive of note arrangement is linear in 1-D, in which ordered ascending pitches in one direction correspond to increasing frequencies. Euler represented harmonic relationships between notes with a mathematical lattice named Tonnetz, which extends the 1-D arrangement into 2-D space by reflecting consonances. Since then, mathematicians, musicians, and psychologists have studied this topic for hundreds of years. Recently, pitch-space modelling has expanded to mapping musical notes into higher-dimensional spaces. This paper aims to investigate existing tonal pitch space models, and to explore a new approach of building a pitch hyperspace by using the Hopf fibration.

Keywords: Tonal Pitch Space · Hopf Fibration · Human Pitch Perception

1 Introduction

Pitch represents the logarithmic value of the fundamental frequency, which is used by humans to distinguish different sounds [10]. The distance between pitches p and q, which has a measurement with the usual metric on $\mathbb{R}$ with an absolute value of the difference, $\mid q - p \mid$, shows the degree of relatedness, with closely related pitches near to each other, and less closely ones far apart. A collection of distances between pitches is denoted as a *scale*, which consists of a number of pitches and their order of the organization [24]. For example, the western 12-tone equal tempered (12-TET) scale, also called the chromatic scale, has twelve pitches. Each pitch has a semitone distance higher or lower than its adjacent one. In the $\mathbb{R}$ Euclidean space, a musical scale could be visualized with a line segment connecting all 12 pitches, since a line segment indicates the range of the collection as well as the increment of all values upon it.

In addition to being logarithmic, human pitch perception is also *periodic* [11]. The distance between one pitch and its double-frequency pitch is called an octave or perfect octave, and this distance can be considered a unit distance in a periodic scale. In different octaves, the denotation of a pitch class – a set of all pitches with a whole number of octaves apart (in 12-TET, a pitch class is presented by a set of points in the quotient space $\mathbb{R}/12\mathbb{Z}$) – is repeatable with the same

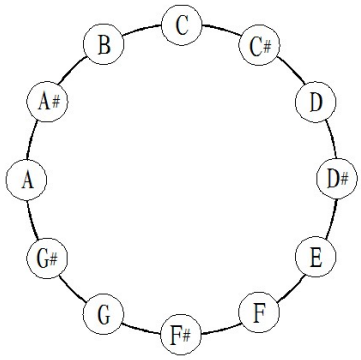


Fig. 1. Circular chroma of a western 12-tone equal tempered scale [13]

symbols or characters but different subscripts, which allows visualization in the $\mathbb{R}^n$ space to show a periodic relationship regardless the hierarchy of tone heights. For example, Fig. 1 shows a *chroma*, an attribute of pitch referring its quality in human pitch-perception, just like hue is an attribute of color, in 1-sphere ($\mathbb{S}^1$ that can be embedded in $\mathbb{R}^2$ Euclidean space) that is the inherent circularity of the chromatic scale in an arbitrary octave. A 12-TET chromatic scale can begin on any note, and proceed through all others before returning to the same note, one octave further in tone height.

In spite of the fact that variations of frequencies describe the interval – a physical distance between pitches – some special collections of pitches or pitch-class sets show "closer" distance (pitches sound more consonant) in perception [9]. For instance, a pair of pitches that have the Perfect fifth relationship sounds more harmonious than a pair of two adjacent pitches in chromatic scales, even though the frequency distance between the notes of a Perfect fifth is much larger than the frequency distance between two adjacent pitches [20]. The physical cause of this perception of consonance relates to the alignment of harmonics in the harmonic series. In a Perfect fifth, every second harmonic of the higher note aligns closely with every third harmonic of the lower note, since the frequency ratio of the two notes is close to 3:2. Dissonant intervals have larger denominator ratios (e.g. 16:15 for a just-tuned semitone, $1 : 2^{12}$ for an equal tempered semitone), with fewer aligned harmonics and more harmonics interfering with each other. Thus, intervals with frequencies close to a small whole-number ratio contain pairs of notes that are somehow "closer together" in terms of perceptual harmonic relationships. In this way, we can imagine a collection of "distance" metrics besides linear frequency distance, in an attempt to indicate the harmonic or musical "closeness" of a pair of notes, relating to the interval between them rather than the distance between them.

Since any two non-identical intervals are independent to each other, the perceptual distance of harmonic closeness can be represented by orthogonal vectors representing the intervals between them. Mathematically, the selective combi-

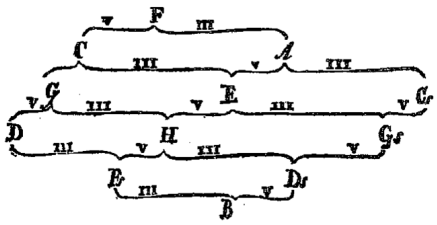


Fig. 2. Euler's Tonnetz (1737) [5]

nation of these vectors can be bundled into groups. To visualize these groups, researchers have modeled the relationships between notes using graphs, tilings, lattices and geometrical figures such as helices. All models like these, which may be multidimensional, are named *pitch spaces* or *tonal pitch spaces* [13]. To model an n-dimensional pitch space, an $\mathbb{R}^n$ pitch space is needed due to the orthogonality of the groups whose order (the number of elements in its set) is n. Each relationship between a set of pitches is represented as a vector in $\mathbb{R}^n$ pitch space, and the distance along that vector corresponds to an accumulated number of the indicated interval (a distance of 3 Minor thirds, for example).

Tonal pitch space research relating to groups in $\mathbb{R}^n$ lies at the intersection of two seemingly independent disciplines: human pitch perception and algebra topology. Human pitch perception is a part of psychoacoustics that studies the psychological and physiological responses associated with sound. Algebra topology leverages tools of abstract algebra to study topological spaces. In this paper, existing tonal pitch space lattices and human pitch perception models will be reviewed in section 2 to section 4. An exploration of modelling tonal pitch spaces in $\mathbb{R}^4$ space with Hopf fibration, which associates with all the lattices and models mentioned in previous sections, will be described in section 5, before the conclusion in section 6.

2 Tonnetz in $\mathbb{R}^2$ space and its Isomorphism

A Swiss mathematician, Leonhard Euler, introduced a conceptual $\mathbb{R}^2$ lattice diagram named Tonnetz (Tone-network in German) representing 2-dimensional tonal pitch spaces in 1737 [5]. Euler's Tonnetz (Fig. 2) shows the triadic relationships of the Perfect fifth and the Major third. Proceeding from the top of the figure, between note F and note C, there is a Perfect fifth (marked as Roman numeral "V"), while from F to A there is a Major third (marked as "III").

In 1858, the $\mathbb{R}^2$ pitch space was rediscovered by Ernst Naumann. Later in 1866, it was disseminated in Arthur von Oettingen and Hugo Riemann's publication of exploration to chart harmonic motion between chords [4]. Oettingen and Riemann show that the relationships in the chart can be extended through *Just intonation* (requiring strict whole number frequency ratios between notes) to form a never-ending sequence in every direction without repeating any

pitches [21]. Modern music theorists generally construct the Tonnetz with an equilateral triangular lattice based on Neo–Riemannian theory. As shown in Fig. 3, the equilateral triangular lattice demonstrates the equal temperament of triads since a minor third (*C*–*E*♭) followed by a major third (*E*♭–*G*) is equivalent to a Perfect fifth (*C*–*G*)

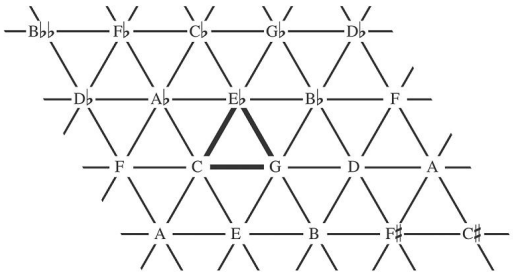


Fig. 3. Modern rendering of the Tonnetz with equilateral triangular lattice [21]

The musical relationships in the lattice can be explained by using group and set theory. The collection of notes along each direction can be considered as a *subgroup* (a subset of a group which is also a group) of a chromatic scale denoted as $\mathbb{R}/12\mathbb{Z}$ under the operation of addition denoted as $+$. For example, the Major third relationship exists as four subsets of the complete chromatic scale: $\{0, 4, 8\}$, $\{1, 5, 9\}$, $\{2, 6, 10\}$ and $\{3, 7, 11\}$. In each of these instances, a collection of notes selected from a chromatic scale are all related as major thirds. for example, for the chromatic scale starting at *C*, the subset $\{0, 4, 8\}$ corresponds to the notes *C, E, A*♭, which appear aligned in Fig. 3 (note than in 12-TET, *F*♭ is an enharmonic spelling of the note *E* and thus equivalent in pitch).

Similarly, the Minor third relationship exists as three subsets of the chromatic scale, as $\{0, 3, 6, 9\}$, $\{1, 4, 7, 10\}$ and $\{2, 5, 8, 11\}$. The Perfect fifth has the same collection of notes with chromatic scale but in different order of arrangement, which is $\{0, 7, 2, 9, 4, 11, 6, 1, 8, 3, 10, 5\}$, and this is also considered as a subgroup [6,17]. In this way, although only some notes can be reached by any combination of jumps of Major Thirds, all notes can be reached by some number of jumps of Perfect Fifths.

In abstract algebra, a group isomorphism presents a one-to-one correspondence between two groups when a group operation is given. The general definition of group isomorphism is:

Considering two groups G and H, where G is under the operation of $\odot$ *and H is under the operation of* $\Diamond$*, where* $\odot$ *and* $\Diamond$ *can only be the operations of addition (denoted as* $+$*) or multiplication (denoted as* $*$*). G and H are isomorphic when there exists a function* $f : G \longmapsto H$ *that fulfills the equation* $f(x)\Diamond f(y) = f(x \odot y)$ *where* $x, y \in G$*, and also they are bijection [1].*

where bijection corresponds to a one-to-one mapping from one group to the other. Applied to tonal pitch spaces, the additive group is sufficient for the consideration of isomorphism. It is easy to show that two subgroups of $\mathbb{R}/12\mathbb{Z}$ are isomorphic under the operation of $+$ in a certain collection of relationships such as Major third, Minor third and Perfect fifth. For example, there are two subgroups under the operation of $+$ from Major third: the subgroup A:$\{0, 4, 8\}$ and the subgroup B:$\{1, 5, 9\}$. A and B are isomorphic if two requirements are fulfilled:

1. A function $f(x)$ can be found, which fulfills the equation
$$f(\{0, 4, 8\}) + f(\{1, 5, 9\}) = f(\{0, 4, 8\} + \{1, 5, 9\})$$
2. The two subgroups are a bijection.

The Major third relationship, which spans four semitones, requires the target function $f(x) = x \pm 4$ to generate the subgroup. In addition, the elements from $\{0, 4, 8\}$ and the elements from $\{1, 5, 9\}$ are a bijection (one-on-one correspondence) as $0 \mapsto 1$, $4 \mapsto 5$, and $8 \mapsto 9$ respectively. Therefore, these two subgroups are isomorphic.

This group isomorphism is directly manifest in music theory as *transposition invariance* (where changing the degree of the starting note of a musical construct does not change the perception of the musical construct, notwithstanding perfect pitch perception) and *tuning invariance* (where changing the frequency of the starting note of a musical construct does not change the perception of the musical construct, notwithstanding perfect pitch perception) [15]. Moreover, the isomorphism of the equilateral triangular lattice in Fig. 3 is useful of building musical keyboards since the dual of equilateral triangle is the hexagon which is a regular polygon of the dihedral group that is a unified tile capable of implementing transpositional invariance [8].

3 Shepard's Double Helix model in $\mathbb{R}^3$ and its winding Torus in $\mathbb{R}^4$

In addition to mathematicians and musicians trying to chart the relationships of pitches, psychologists also explore ways to depict the psychological responses of musical sounds. In human pitch perception, Shepard introduced a double helix model of visualizing ascending chroma and Perfect fifth relationship simultaneously. [22]

Fig. 4 (a) shows an equilateral triangular lattice in $\mathbb{R}^2$ space. The arrangement in vertical direction represents the whole-tone relationship such as C to D, and the zigzag between the two vertical edges depicts the semi-tone relationship such as C to $C\sharp$. As shown in Fig.4 (b), when twisting the lattice along the vertical direction which indicates the ascending (or descending) of octave, the lattice of equilateral triangles form a double helix structure. If this double helix structure is mapped onto a cylinder in $\mathbb{R}^3$ space, the resulting cylindrical lattice is as shown in Fig.4 (c).

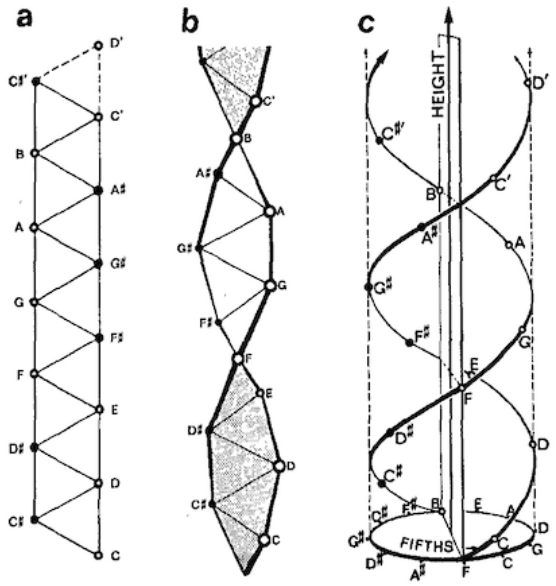


Fig. 4. Shepard's Double Helix model of human pitch perception. [22] (a) the flat strip of equilateral triangles in $\mathbb{R}^2$; (b) the strip of triangles given one complete twist per octave; (c) the resulting double helix shown as a winding around a cylinder in $\mathbb{R}^3$.

The cylindrical double helix structure can collapse vertically so as to continuously vary the structure that representing the pitches between the double helix with the rectilinear axis in Fig. 4 (c), and eventually becomes the variant with a completely circular axis in Fig. 5. This new structure is a torus in $\mathbb{R}^4$ space, and it embeds the cylindrical lattice. In addition, the pair of edges in Fig. 4 (a) becomes two circles linked together in the torus, which actually is a "Hopf link" [12]. In Fig. 5, one circle consists of notes *C, D, E, F*♯*, G*♯ and *A*♯, and the other circle includes notes *C*♯*, D*♯*, F, G, A* and *B* respectively.

4 Multidimensional Pitch Space Torus and Phases of Pitch-Class Sets

In the last two decades, the exploration of modelling higher dimensional tonal pitch space shown solid results, such as Dmitri Tymoczko in 2006 exploited non-Euclidean geometry space to represent a musical chord, and pointed out that the structure is an n-orbifold [23]. In 2013, Gilles Baroin introduced a model in $\mathbb{R}^4$ space which derived from two planar graphs: Triangle (C3) and Square (C4). There are two axes in the model which indicate the rotation directions. The two axes in Fig. 6 are represented by straight lines. However, if a straight line was considered as the special case of an infinite circle $\mathbb{S}^1$ mapped into a $\mathbb{R}$ space, the Cartesian product of these two perpendicular circles is represented by a torus

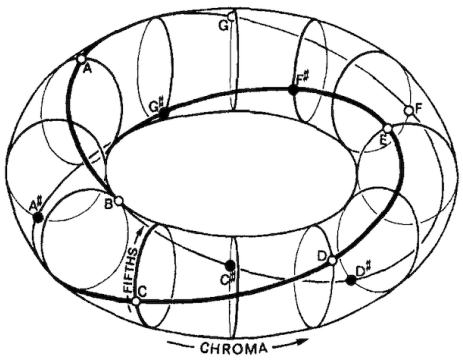


Fig. 5. Double Helix winding a torus in $\mathbb{R}^4$ [22]

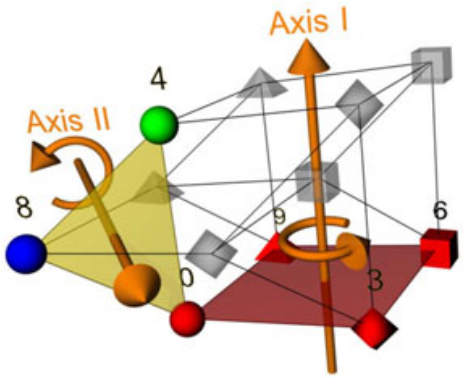


Fig. 6. Major Axes of the 4D model [3]

($\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$) in $\mathbb{R}^4$ space. Conclusively, it turns out the traditional Tonnetz can be embedded into a 4-dimensional hypersphere [3].

Besides leveraging algebraic topological structures to extend dimensions, another approach is mapping pitches from real space into complex space. In 2012 Emmanuel Amiot decomposed the phases of the pitch-class sets with Fourier coefficients, which delivered an even finer result on mapping chromatic scale notes onto a 4-dimensional torus surface [2].

5 Villarceau Circles and Hopf Fibration

Before introducing Villarceau Circles and Hopf Fibration, it is necessary to have a review of the 4-dimensional torus $\mathbb{T}^2$ as used in human pitch perception models.

Since a pitch class from the 12-tone equal tempered scale is presented by a set of points in the quotient space $\mathbb{R}/12\mathbb{Z}$, n pitch classes can be denoted as $\mathbb{R}^n/12\mathbb{Z}^n$. Hence, an n-torus can be represented by the equation $\mathbb{T}^n = \mathbb{R}^n/12\mathbb{Z}^n$. Whereas a chromatic scale is represented by a circle which is denoted as $\mathbb{S}^1$ in algebraic topology, there is the other equation $\mathbb{S}^1 = \mathbb{R}/12\mathbb{Z}$. Combining two equations together, it is easy to have $\mathbb{T}^n = (\mathbb{R}/12\mathbb{Z})^n = \left(\mathbb{S}^1\right)^n$. When n equals 2, the equation becomes $\mathbb{T}^2 = \left(\mathbb{S}^1\right)^2 = \mathbb{S}^1 \times \mathbb{S}^1$. This indicates a torus in $\mathbb{R}^4$ space could be represented by the Cartesian product of two circles $\mathbb{S}^1$ in $\mathbb{R}^2$ space. One of the circles is around its axis of rotational symmetry and the other one is around a circle in the interior of the torus.

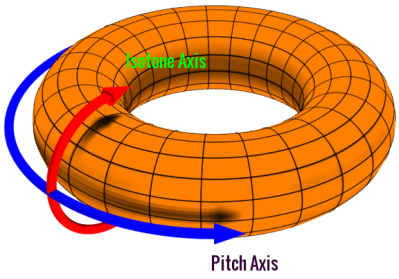


Fig. 7. Two axes for mapping Neo-Riemann Tonnetz

Comparing to the torus in Fig. 5, the toroidal model makes sense in human pitch perception when the circle around its axis of rotational symmetry (or a.k.a. along "Toroidal" direction) represents chroma, and the circle around a circle in the interior of the torus (along the "Poloidal" direction) represent the

repetition of pitches. According to [19], the repetition of the pitch direction is named "Isotone Axis", and the ascending chroma direction is named "Pitch axis" respectively. The torus in Fig. 7 can be used to map the Neo-Riemann Tonnetz onto the surface. It should be noted that this arrangement (chroma / pitch in the torodial direction, isotone in the polodial direction) may be reversed without losing generality of the model, but most researchers have used this first arrangement and we maintain that approach.

In geometry, cutting a torus with a plane bitangent to two symmetrical poloidal circles results in a pair of *Villarceau* circles. This can be seen in Fig.8 where the bitangent plane ε is shown in pink and the coplanar pair of Villarceau circles ($M1$ and $M2$, marked in red and blue) are produced by this cutting. These two Villarceau circles are linked to each other [16]. Since the torus has symmetry of its centre, a pair of mirrored villarceau circles can easily be generated by rotating the torus along toroidal direction with 180 degrees. Each Villarceau circle and its 180-degree mirrored circle are also linked together. This link is a Hopf link, and the circle of a Hopf link is a Hopf fiber [14].

The *Hopf Fibration* (also known as the Hopf bundle or Hopf map), named after German geometer and topologist Heinz Hopf [7] is a foundation stone in the theory of Lie Groups. The Hopf Fibration describes a hypersphere (called a 3-sphere in $\mathbb{R}^4$ space) in terms of circles (the "fiber") and an ordinary sphere [14].

The denotation of fiber bundle (bundle of linked circles) is:

$$\mathbb{S}^1 \hookrightarrow \mathbb{S}^3 \overset{p}{\mapsto} \mathbb{S}^2$$

which means a circle $\mathbb{S}^1$ is embedded in the hypersphere $\mathbb{S}^3$, and the Hopf map $p : \mathbb{S}^3 \mapsto \mathbb{S}^2$ projects $\mathbb{S}^3$ onto an ordinary 2-sphere $\mathbb{S}^2$.

Theoretically, an $\mathbb{R}^3$ space can be filled with nested tori made of linking Villarceau Circles, which is induced by Stereographic projection of the Hopf fibration [14]. As shown in Fig. 9, the same type of Villarceau Circles (or Hopf fibers) are considered "parallel".

Because these fibers are parallel, we can apply any of the original pitch-class mappings onto this fibration, and specifically, the New-Riemann Tonnetz can be mapped onto a torus. Along each direction of Fig.3's equilateral triangular lattice, the subgroups representing a relationship of pitches are always parallel. Therefore, to model a 4-dimensional tonal pitch space with Hopf fibers, two subgroups of the pitches (in $\mathbb{R}/12\mathbb{Z}$ space) need to be selected to map onto the a Hopf link (including two types of Hopf fibers). For example, the Perfect fifth is mapped onto one type of Hopf fiber, and the Major third is mapped onto the other type of Hopf fiber, where two Hopf fibers combine to make a Hopf link.

However, the poloidal direction of the torus may not always show the partition of same pitch in a single octave. Instead, it could show the partition of the pitch that is several octaves when the poloidal circle of the torus gets larger. To force the poloidal circle to show a relationship of pitches in a single octave, the paralleled Villarceau Circles have to vary to non-round closed curves but still keep parallel structure. For example, in Fig. 10, the Perfect fifth (in blue) and the Major third (in red) represent two relationship of pitches individually,

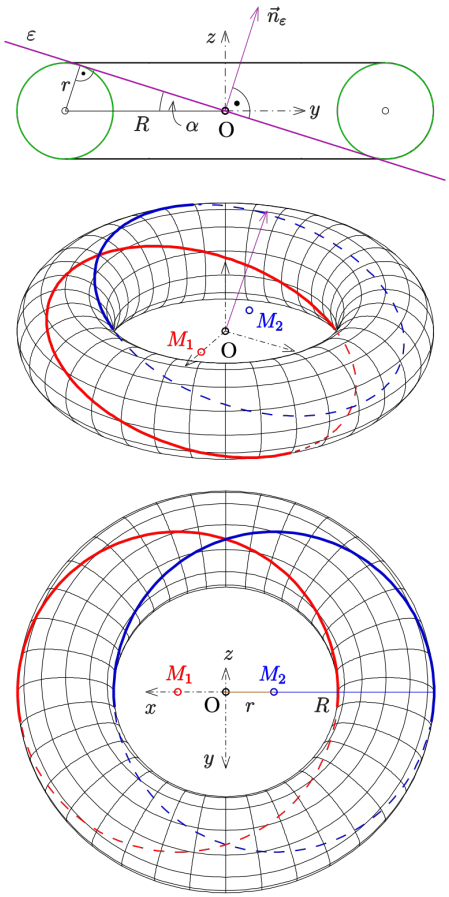


Fig. 8. Villarceau circles in a Torus [25]

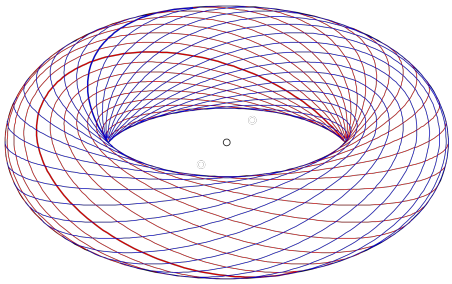


Fig. 9. Torus with two types (original and mirrored) of Villarceau circles [25]

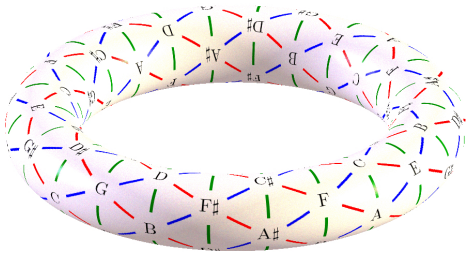


Fig. 10. One toroidal view of the neo-Riemannian Tonnetz [26]

the Minor third (in green) has been laid along the poloidal direction. Because the Minor third poloidal circle represents a subgroup of pitches in $\mathbb{R}/12\mathbb{Z}$ space rather than Minor third pitches with partition in $\mathbb{R}$ space, it is obvious that the paralleled red curves and blue curves are not, in fact, circles anymore.

6 Conclusion and future work

In this paper, the existing human pitch perception models and tonal pitch spaces lattices are well studied. Through the line segments in $\mathbb{R}$ space, the circular chromatic scale in $\mathbb{R}^2$ space, the double helix model in $\mathbb{R}^3$ space, and the torus in $\mathbb{R}^4$ space, we attempted to find a generic way of modelling the tonal pitch space onto manifolds. After exploring the approaches in the interaction of two seemingly independent disciplines human pitch perception and algebra topology, a new method of modelling 4-dimensional tonal pitch spaces is presented, which leverages the knowledge of Hopf fibration.

In the future, one possible research could be the torus knots which presenting the sinusoidal phases of the musical notes on a Hopf fibers bundled torus. Topologically, a knot is a continuous looped path that may be intertwined with itself in one or more ways. When a knot lies on the surface of a torus in $\mathbb{R}^3$ space, it is called torus knot. This torus knot is denoted with two numbers: (p, q), where p indicates the number of times the knot's path goes around its axis of rotational symmetry, and q indicates the number of times the knot's path goes around a circle in the interior of the torus [18]. For example, Fig. 11 shows a torus knot that could be applied to a Neo-Riemann Tonnetz mapping torus. The 12 black dots indicate the 12 notes within a chromatic scale with different phases in complex space [2]. It is easy to count the number of rounding the rotational symmetrical axis and the number of rounding circle of interior of the torus. Apparently, in this example, this torus knot is a (3, 4) knot.

Knots can represent the path of a subgroup different from the subgroups used to construct the original tonnetz mapping. In other words, the different combinations of p and q, in the example, reflect how a group of chromatic scale can be divided into subgroups. No matter which combination is used, in order to map the chromatic scale which has 12 notes, the torus knot (p, q) would need to fulfil the equation $p \cdot q = 12$. Though the mathematical model for the

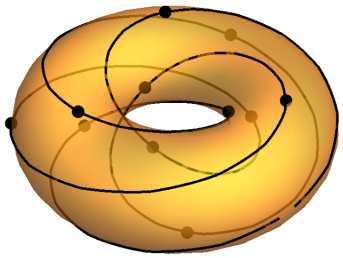


Fig. 11. Torus knot (3,4) shows the continuous orbits of all pitches of the chromatic scale in a $\mathbb{T}^2$ Torus [3]

torus knots can easily be created, more exploration and research are needed to study the position of musical notes on the surface of the knot, which potentially presents the sinusoidal phases of the musical notes. In addition, the torus knots can be further utilized in the research of human pitch perception.

References

1. Allenby, R.: Rings, Fields and Groups, An Introduction to Abstract Algebra. Butterworth-Heinemann (1991)
2. Amiot, E.: The torii of phases. In: Yust, J., Wild, J., Burgoyne, J.A. (eds.) Mathematics and Computation in Music. pp. 1–18. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
3. Baroin, G.: The planet-4d model: An original hypersymmetric music space based on graph theory. In: Proceedings of the Third International Conference on Mathematics and Computation in Music. pp. 326–329. MCM 2011, Springer-Verlag, Berlin, Heidelberg (2011)
4. Cohn, R.: Introduction to neo-riemannian theory: A survey and a historical perspective. Journal of Music Theory 42(2), pp. 167–180 (1998)
5. Euler, L.: De harmoniae veris principiis per speculum musicum repraesentatis. In: Novi commentarii academiae scientiarum Petropolitanae. pp. 330—353. St. Petersburg (1774)
6. Fiore, T.M., Noll, T.: Commuting groups and the topos of triads. In: Proceedings of Mathematics and Computation in Music - Third International Conference, MCM 2011. pp. 69–83. Paris, France (2011)
7. Hopf, H.: Über die abbildungen der dreidimensionalen sphäre auf die kugelfläche. Mathematische Annalen 104(1), 637–665 (1931)
8. Hu, H., Gerhard, D.: WebHexIso: A Customizable Web-based Hexagonal Isomorphic Musical Keyboard Interface. In: Proceedings of the 42th. International Computer Music Conference. pp. 294—297. Utrecht, Netherlands (2016)
9. Hu, H., Park, B., Gerhard, D.: Mapping Tone Helixes to Cylindrical Lattices Using Chiral Angles. In: Proceedings of the 12th International Sound and Music Computing Conference. pp. 447—454. Maynooth, Ireland (2015)
10. Klapuri, A., Davy, M.: Signal Processing Methods for Music Transcription. Springer-Verlag, Berlin, Heidelberg (2006)

11. Krumhansl, C.L.: Perceptual Structures for Tonal Music. Music Perception pp. 28—62 (1983)
12. Kusner, R.B., Sullivan, J.M.: On distortion and thickness of knots (1997)
13. Lerdahl, F.: Tonal Pitch Space. Oxford Univ. Press (2001)
14. Lyons, D.W.: An elementary introduction to the hopf fibration. Mathematics Magazine 76(2), 87–98 (2003), http://www.jstor.org/stable/3219300
15. Milne, A., Sethares, W., Plamondon, J.: Tuning continua and keyboard layouts pp. 1—19 (2008)
16. Monera, M.G., Monterde, J.: Building a Torus with Villarceau Sections. Journal for Geometry and Graphics 15(1), 93–99 (2011)
17. Morris, R.D.: John Rahn. Basic Atonal Theory. New York: Longman, 1980. Music Theory Spectrum 4(1), 138–154 (1982), https://doi.org/10.2307/746016
18. Murasugi, K.: Knot Theory and Its Applications. Birkhäuser Boston (2008)
19. Park, B., Gerhard, D.: Discrete Isomorphic Completeness and a Unified Isomorphic Layout Format. In: Proceedings of the Sound and Music Computing Conference. Stockholm, Sweden (2013)
20. Piston, W., DeVoto, M.: Harmony. W. W. Norton & Company (1987)
21. Riemann, H.: Ideen zu einer Lehre von den Tonvorstellungen, Jahrbuch der Bibliothek pp. 21—22 (1914—1915)
22. Shepard, R.N.: Geometrical Approximations to the Structure of Musical Pitch. Psychological Review 89(4) (1982)
23. Tymoczko, D.: The geometry of musical chords. Science 313(5783), 72–74 (2006), https://science.sciencemag.org/content/313/5783/72
24. Tymoczko, D.: Three conceptions of musical distance. In: Mathematics and Computation in Music. pp. 258–272. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
25. Wikipedia contributors: Villarceau circles — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Villarceau_circles&oldid=875134100 (2018), [Online; accessed 10-July-2019]
26. Wikipedia contributors: Neo-riemannian theory — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Neo-Riemannian_theory&oldid=893433197 (2019), [Online; accessed 10-July-2019]

Automatic Dastgah Recognition Using Markov Models

Luciano Ciamarone[1], Baris Bozkurt[2], and Xavier Serra[2]

[1] Independent researcher
[2] Universitat Pompeu Fabra, Barcelona, Spain
luciano.ciamarone@libero.it, baris.bozkurt@upf.edu, xavier.serra@upf.edu

Abstract. This work focuses on automatic Dastgah recognition of monophonic audio recordings of Iranian music using Markov Models. We present an automatic recognition system that models the sequence of intervals computed from quantized pitch data (estimated from audio) with Markov processes. Classification of an audio file is performed by finding the closest match between the Markov matrix of the file and the (template) matrices computed from the database for each Dastgah. Applying a leave-one-out evaluation strategy on a dataset comprised of 73 files, an accuracy of 0.986 has been observed for one of the four tested distance calculation methods.

Keywords: mode recognition, dastgah recognition, iranian music

1 Introduction

The presented study represents the first attempt in applying Markov Model to a non-western musical mode recognition task. The proposed approach focuses on Persian musical modes which are called *Dastgah*. Several different approaches to the same task have already been documented. In 2011 Abdoli [3] achieved an overall accuracy of 0.85 on a 5 Dastgahs classification task by computing similarity measures between Interval Type 2 Fuzzy Sets. In 2016 Heydarian [5] compared the performances of different methods including chroma features, spectral average, pitch histograms and the use of symbolic data. He reported an accuracy of 0.9 using spectral average and Manhattan metric as a distance calculation method between a signal and a set of templates. Another contribution to the field comes from research works in Turkish music [11] and [4] from which this work inherit some analysis techniques [8]. None of these previous works use Markov models for their classification purposes while for western music several applications has been explored [12][13] although only for chord progression thus under the point of view of music harmony. The presented work, instead, investigates the music melodic aspect developing a mathematical model able to encode the typical melodic interval progression of Persian Dastgahs in the form of Markov Transition Probabilities Matrices (section 3.2). A subsequent distance evaluation method between matrices has been applied in order to carry out the classification task (section 3.3). Finally, standard machine learning evaluation

has been carried to measure system performances. An accuracy of 0.98 has been reached on a database of 73 audio files belonging to the seven main Persian Dastgahs. The complete algorithm has been publicly shared on a github repository[7] for the sake of reproducibility. In the final part of the presented paper future developments of this research have been identified.

2 Persian Traditional Music

Persian music is based on a set of seven principal Dastgahs: *shur*, *homayun*, *segah*, *chahargah*, *mahour*, *rast-panjgah* and *nava*. The seven main modes and their five derivatives (*abu ata*, *bayat-e tork*, *afshari*, *dashti* and *bayat-e esfehan*) are collectively called the twelve dastgahs, they cover most of the structures in Persian music [3] [5].

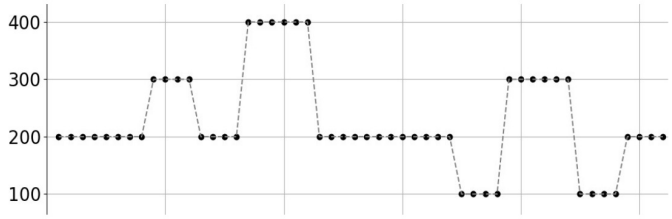
Traditional Persian music has the octave concept and one octave always contains seven principal notes. The tuning system of Persian music applies 24 quarter tones per octave[2] and does not rely on equal temperament. The most authentic definition has been given by Farhat [1] who teaches that Persian music has very characteristic intervals, one of them is the neutral second. This is a flexible interval but in all its variations it is noticeably larger then the minor second (western semitone) and smaller then the major second (western whole tone). Another typical interval [1] is an interval which is larger then the major second (western whole tone) but smaller then the western minor third. Rhythmic structure of Persian music is generally strictly connected to voice and speech, often music is conceived as accompaniment for singers. Each *Dastgah* consists of some partial melodies called *Gushe*, the arrangement of *Gushe*s during the performance is called *Radif*. Conceptually Persian music is conceived like melodic motives around a central tone and modulation is conceived as changing the central tone.

3 Methodology

The presented Dastgah recognition system applies a sequence of three processes: pitch estimation and quantization, Markov Modeling, and classification. Markov chains have been used for modeling the sequence of musical intervals. The concept behind Persian musical intervals is the same as in western music, which is a frequency ratio between two notes. For example the western musical fifth is defined as $\sqrt[12]{2^7} \cong 3/2$. [2] The presented Markov algorithm models in the form of Transition Probability Matrix, the sequence (in time) of musical intervals contained in each audio file. The synthetic example in Fig. 1 can help understanding the Markov Matrix building strategy. Frequency values in time like in Fig. 2 are converted in a vector of consecutive musical intervals; in this example [*3/2, 3/2, 2, 2, 2, 3, 3, 2*]. Thus we have three Markov states: *A=3/2*, *B=2* and *C=3*.

[2] an example of western fifth is the interval between A4=440Hz and E5=659.25Hz (659.25/440=1.498 $\cong \sqrt[12]{2^7}$

Finally, transitions between consecutive states are counted and cumulated to build up the Markov matrix. As we can see transitions from B to A never occurs (neither C to A and A to C), furthermore if we had a state *D=5/2* it would have occurrence count equal to *0* in this example.



	A	B	C
A	1	0	0
B	1	2	1
C	0	1	1

Fig. 1. Example of Markov Matrix building strategy starting from quantized pitch data (this is not real data)

3.1 Pitch Detection and Quantization

For pitch detection, we have used the algorithm explained in [10] (more specifically the implementation by its authors[3]) which mainly applies post-filters to the output of the Melodia algorithm [9]. According to the implementation[3] of [10], pitch quantization is achieved by mapping each pitch value to the closest pitch histogram peak where pitch histograms are computed as in [8]. Frequency quantization has the effect of stabilizing the values given by the pith detection algorithm forcing them to be equal to the closest pitch histogram peak [8] (for example the sequence of pitches: [200.1, 200.2, 199.85] is forced to be [200, 200, 200] if 200Hz is the closest pitch histogram peak).
Secondly, spurious and very short duration notes has been removed because they are assumed to be detection errors. When a quantized note is removed (because of too short in duration) it is replaced (and merged) with the following stable note (a stable note is assumed to be at least 15 milliseconds long). The quantized pitch data obtained as a result of these two post processing steps are exemplified in Fig. 2 .

3.2 Markov Model

The Markov Model builder block uses the data generated by the pitch detection and quantization algorithm in order to create a vector of musical intervals which are calculated from the ratio between two consecutive frequency values (as explained at the beginning of this *section 3* and in Fig. 1). The resulting sequence of musical intervals (from one audio file) can be concatenated (using

[3] `https://github.com/MTG/predominantmelodymakam`

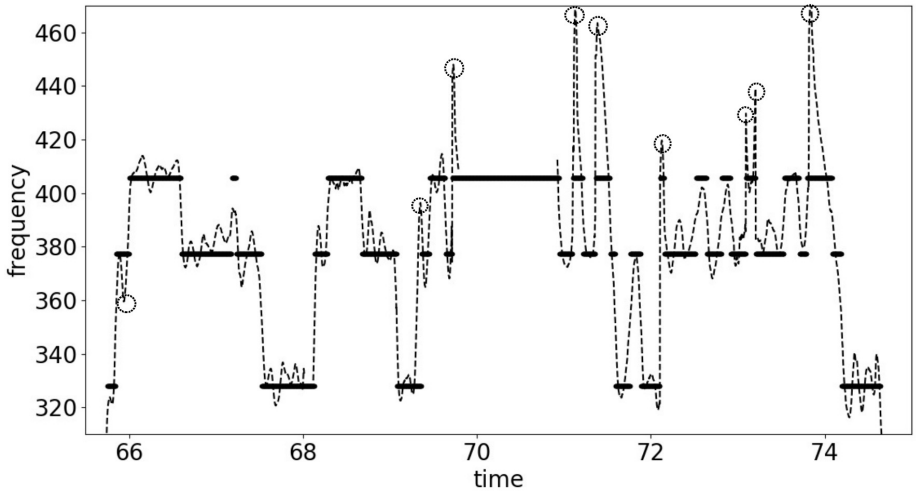


Fig. 2. Ten seconds of a Segah audio sample. *Dashed Line:* Melody profile of the audio input as analyzed by the Essentia monophonic pitch detection algorithm. *Continuous Line:* Same melody profile after post processing for stabilizing the frequency values. *Dotted Circles:* Peaks that generated too short duration quantized notes which have been replaced and merged with the following stable note.

several audio files) in case one wants to build up the database matrix associated with one of the seven possible Dastgah considered in the present work.

Building the Transition Probability Matrix associated with the sequence of musical intervals consists in counting the number of transition occurrences between consecutive intervals (as exemplified in Fig. 1). All probability values have been normalized to one in the end. After that a vector containing the count of the occurrences of each Markov Model state has been created, the Fig. 3 shows this vector calculated for a small audio chunk belonging to Dastgah-e Segah.

The vector plotted in *Fig. 3* is a sort of one-dimensional view of a Markov Matrix, on the x axis there are the Markov Matrix indexes from one to ns (where ns is the total number of Markov States); on y axis there is the normalized frequency of occurrence of each state. The presented Dastgah Recognition system does not use a previously defined TET musical system and there are no octave folding operations. The dimension of the Markov Matrix just needs to be greater then the maximum number of playable musical intervals and sufficiently small not to overload the computational cost of the whole algorithm. Results presented in this paper have been obtained using 98 Markov States (ns=98). Each Markov State represents a bin whose width progresses exponentially with a base of $c = \left(\frac{1}{ns} + 1\right)$ [4], the computed musical intervals are mapped into these bins (this

[4] using 98 Markov States, $c = \left(\frac{1}{98} + 1\right) = 1.0102$ which is smaller then the western music semitone $\sqrt[12]{2} = 1.0595$

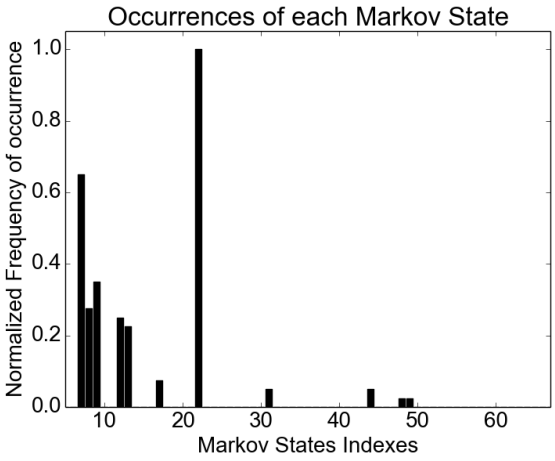


Fig. 3. Markov states frequency of occurrence for a small chunk of Dastgah-e Segah

means that small fluctuations of frequency ratios are mapped into the same bin if those fluctuations are contained in the bin width)[7]. In *Fig. 3* there are lots of Markov states with number of occurrences equal to zero, this means that for that audio sample, no musical intervals have been mapped into those bins, this is why this vector can be considered as a first raw fingerprint of the Dastgah to which the audio file belongs to.

3.3 Classification

In order to classify one unknown audio file as belonging to one of the seven considered families of Dastgah it is necessary to implement a metric able to measure the distance between the matrix associated to the unknown audio file and the seven database matrices associated to the seven Dastgah. In this work four distance candidates have been tested: Euclidean distance (*equation 1*), Kullback-Leibler distance (*equation 2*), Bhattacharyya likelihood (*equation 3*) and the last metric (which is also in the form of a likelihood) State Probability Correlation (*equation 4*). This last metric basically performs the dot product between *Frequency of Occurrence* vectors like the one showed in *Fig. 3*.

$$euclidean = \sqrt{\sum_{i=1}^{ns}\sum_{j=1}^{ns}(\chi_{ij} - db_{ij})^2}\ . \tag{1}$$

$$kullback = \sum_{i=1}^{ns}\sum_{j=1}^{ns}\chi_{ij}\cdot log\left(\frac{\chi_{ij}}{db_{ij}}\right)\ . \tag{2}$$

$$battacharyya = \sum_{i=1}^{ns}\sum_{j=1}^{ns}\sqrt{\chi_{ij}\cdot db_{ij}}\ . \tag{3}$$

$$SPC = \sum_{n=1}^{ns} \bar{\chi_n} \cdot \bar{db_n} \ . \tag{4}$$

Where ns is the number of Markov states, χ_{ij} is one element (scalar) of the unknown matrix, db_{ij} is one element (scalar) of the database matrix; $\bar{\chi_n}$ and $\bar{db_n}$ are the cumulated values (scalars) along the rows of the matrices in order to obtain a cumulative value of occurrence probability for each Markov state.

4 Experiments and Results

The experiments carried out had the goal of testing and validating the classification algorithm, a standard machine learning evaluation procedure has been applied. A *Leave One Out* testing strategy has been implemented and in the end, standard machine learning evaluation parameters have been calculated. A github repository [7] has been created where the testing package can be downloaded and executed again for obtaining the same results presented in this paper.

4.1 Dataset

In the aforementioned github repository an *annotation.json* file can be found. In this file is contained the formal description of the used database of audio files. A total amount of 73 audio files has been collected with average duration of about two minutes each (maximum 4:50, minimum 1:10), they belong to the seven Dastgah category considered in this work: Segah (16 files), RastPanjgah (5 files), Nava (7 files), Mahour (15 files), Homayun (10 files), Chahargah (8 files) and Shur (12 files). The recordings are monophonic (recording of a single instrument) and does not include mode transitions/modulations. No constraint has been applied in selection of instruments. The dataset contains recordings of singing, tar, setar, santur among other instruments.

4.2 Testing Strategy and Results

A *Leave One Out* validation procedure has been considered for the presented work. The procedure uses one sample for testing and all the rest for modeling. This is repeated for each sample and the results are averaged. At the end of validation process the following standard machine learning evaluation measures has been calculated: Recall, Specificity, Precision, False Positive Rate, False Negative Rate, F1 and Accuracy. The four different distance calculation methods are defined in equations 1, 2, 3 and 4. Table 1 shows the evaluation measures obtained for each distance metric.

The first two methods gave very poor results below the 50% of correct answers. The logarithmic terms in *equation 2* makes this metric particularly suitable for very smooth distributions like Gaussian ones; in our case this metric suffers the big entropy (from a statistical point of view) of the data it is applied to. The last method (SPC) gave an *Accuracy* of 0.507 and a *FPR* (*False*

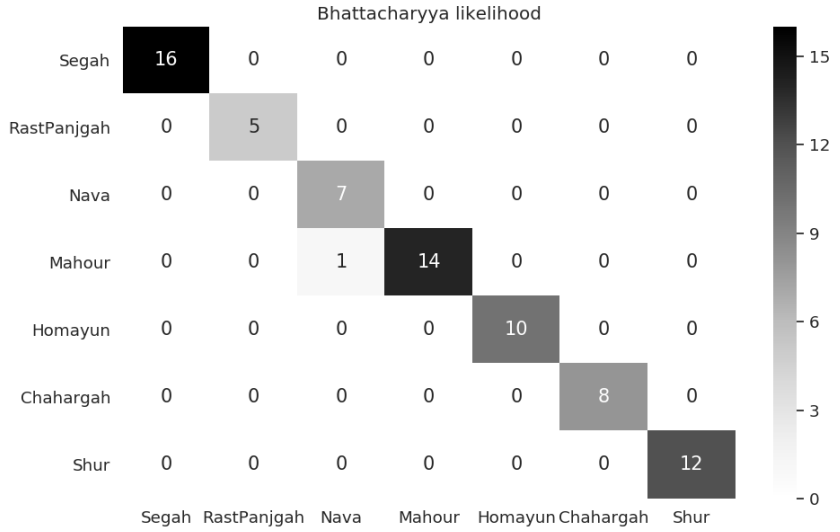


Fig. 4. Confusion matrix for the Bhattacharyya likelihood calculation method

Positive Rate) of 0.493 . The third distance calculation method (*Bhattacharyya likelihood*) gave instead an *Accuracy* of 0.986 and a *FPR* of 0.014 . Fig. 4 shows the confusion matrix for the Bhattacharyya likelihood.

Results clearly state that *Equation 3* is the best way of calculating similarities between Markov Transition Probability Matrices; in fact this method resulted in only one error on 73 audio file which means a percentage of correct answers equal to 98.6%. The only error of the *Bhattacharyya* classification method is an audio file belonging to the *Mahour* family which has been classified as *Nava*. The reason why all the other metrics are so far from the Battacharyya metric in terms of results, is that *equation 3* is the only one which uses an *element by element* product between Markov Matrices, this produces the effect that only values different from zero contribute to the total sum; non-zero values represent specific transitions between consecutive musical intervals and these transitions are the musical core of *Dastgah*s.

metric	recall	specificity	precision	FPR	FNR	F1	accuracy
euclidean	0.061	0.663	0.247	0.753	0.939	0.098	0.247
kullback	0.014	0.306	0.068	0.931	0.986	0.024	0.068
battacharyya	0.935	0.998	0.986	0.014	0.065	0.960	0.986
SPC	0.170	0.860	0.507	0.493	0.829	0.255	0.507

Table 1. Scores

5 Conclusions and Further Developments

In this work we approached the problem of Persian Dastgah recognition and classification using Markov Models. The presented Dastgah recognition system has been tested following a standard machine learning evaluation procedure and it gave a maximum accuracy of 0.986. Results show that Markov Models are able to encode information about the content of each Dastgah in terms of musical intervals and their temporal sequence. The presented system has been tested on monophonic recordings of short duration. Our further efforts will be dedicated to building a larger dataset including longer recordings and improvisations as well as multi instrumental recordings.

References

1. Farhat, H., The dastgh concept in Persian music, Cambridge University Press, Cambridge, 1990.
2. Vaziri, A. N.: *Dastur-e Tar*, Tehran, 1913.
3. Abdoli, S., Iranian traditional music dastgh classification, Proceedings of the 12 th International Conference on Music Information Retrieval (ISMIR), Miami, 2011.
4. B. Bozkurt, R. Ayangil, A. Holzapfel, 2014, "Computational Analysis of Turkish Makam Music: Review of State-of-the-Art and Challenges, Journal of New Music Research, 43:1, pp. 3-23.
5. Heydarian, P., Automatic recognition of Persian musical modes in audio musical signals, PhD thesis, London Metropolitan University, 2016.
6. C++ library for audio and music analysis, description and synthesis, including Python bindings, [`https://github.com/MTG/essentia`].
7. Dastgah-Recognition-System github repository, [`https://github.com/luciamarock/Dastgah-Recognition-System`].
8. A. C. Gedik and B. Bozkurt, Pitch-frequency histogram-based music information retrieval for Turkish music, Signal Processing, vol. 90, no. 4, pp. 10491063, 2010.
9. J. Salamon and E. Gomez, Melody extraction from polyphonic music signals using pitch contour characteristics, in IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 6, 2012, pp. 17591770.
10. H. S. Atli, B. Uyar, S. Senturk, B. Bozkurt, and X. Serra, Audio feature extraction for exploring Turkish makam music, in Proc. of 3rd Int. Conf. on Audio Technologies for Music and Media, Ankara, 2014, pp. 142153.
11. S. Senturk, Computational analysis of audio recordings and music scores for the description and discovery of Ottoman-Turkish makam music, Ph.D. dissertation, Universitat Pompeu Fabra, 2016.
12. Noland, K., Computational tonality estimation: signal processing and Hidden Markov Models, PhD thesis, Queen Mary, University of London, 2009.
13. Noland, K. and Sandler, M., Key estimation using a Hidden Markov Model. Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), Victoria, Canada, 2006.

Chord Function Identification with Modulation Detection Based on HMM

Yui Uehara[1], Eita Nakamura[2], and Satoshi Tojo[1] *

[1] Japan Advanced Institute of Science and Technology
[2] Kyoto University
{yuehara, tojo}@jaist.ac.jp, enakamura@sap.ist.i.kyoto-u.ac.jp

Abstract. This study aims at identifying the chord functions by statistical machine learning. Those functions found in the traditional harmony theory are not versatile for the various music styles, and we envisage that the statistical method would more faithfully reflect the music style we have targeted. In machine learning, we adopt hidden Markov models (HMMs); we evaluate the performance by perplexity and optimize the parameterization of HMM for each given number of hidden states. Thereafter, we apply the acquired parameters to the detection of modulation. We evaluate the plausibility of the partitioning by modulation by the likelihood value and, as our innovative method, the result is reduced back to the number of states conversely. As a result, we found that the six-state model outperformed the other models both for the major keys and for the minor keys although they assigned different functional roles to the two tonalities.

Keywords: chord function; hidden Markov model; modulation detection.

1 Introduction

The chord functions are one of the most fundamental bases of tonal music to identify the key. Although the traditional functional harmony theory well describes the general roles of chords, the functions should have been diversified in accordance with the target music.

Previously chord function identification has been carried out mainly by statistical clustering algorithms [4, 8]. Since these statistical methods learn from raw data instead of the textbook theory, they have the potential to reflect the difference of music styles. A recent study proposed a generative model [12], which is advantageous in its predictive power and in its applicability to practical problems such as melody harmonization [11]. However, this study focused on popular music and the key was assumed to be invariant within each piece. In our research, we consult J. S. Bach's music, thus the modulation detection would be inevitable. Thus far, modulation detection has been carried out either

* This research has been supported by JSPS KAHENHI Nos. 16H01744 and 19K20340.

by heuristics [8] or by a key-finding algorithm [4] though there have been still several difficult cases to determine the key [10].

We conceive that the local keys could be also determined by the functional progression of chords. Therefore, we propose a new dynamic modulation detection method, applying the statistically found chord functions. Here, the optimal number of functions would be also determined computationally so that it maximizes the likelihood of chord progressions in the entire corpus. In this research, we achieve the detection of the data-oriented chord functions, together with the detection of modulation. We envisage that we would obtain finer-grained chord functions which faithfully reflect the targeted music style. Our method is new in that we do not need to prefix the scope of modulation as opposed to the algorithm using the histogram of pitch classes [5, 10, 3, 13].

We begin this paper by reviewing related works, especially the key detection algorithms and the statistical learning methods of the chord functions in section 2. Then, we propose our method in section 3, and thereafter show the experimental results in section 4. We conclude in section 5.

2 Related Work

2.1 Key detection algorithms

Among the key detection algorithms based on the histogram of the pitch classes [10, 5, 3, 13], the most widely used one is the Krumhansl-Schmuckler algorithm that adopts the key-profile obtained by a psychological experiment [5]. More recently, the key-profile was obtained from music data by using a simple Bayesian probabilistic method [10] and the Latent Dirichlet Allocation (LDA) [3].

Sakamoto et al. [9] employed the distance between chords by using Tonal Pitch Space (TPS) [6] rather than the pitch classes. Given a sequence of Berklee chord names, the key is detected by the Viterbi algorithm, not requiring a fixed scope. A Berklee chord can be interpreted in multiple keys, for example, the chord **C** is **I** of C major key as well as **IV** of G major key. Therefore, the network of candidate nodes consists of keys with degree names. Since TPS does not have adjustable parameters, it cannot reflect the difference in music styles.

2.2 Statistical learning of the chord functions

Statistical learning of the chord functions has been studied by classifying the chords using clustering algorithms. Rohrmeier and Cross [8] used the hierarchical cluster analysis to find the statistical properties of the chords, where the most distinctive cluster of the pitch class sets reflected the dominant motion in both major and minor keys. They also found that the result for the minor key was significantly different from that for the major key. The clusters that represent the Tonic and Dominant of the relative major key were obtained.

Jacoby et al. [4] also carried out the clustering of the chords in J. S. Bach's chorales and some other datasets. They proposed the evaluation method using

two criteria, accuracy and complexity, inspired by the information theory. They introduced the optimal complexity-accuracy curve, which is formed by the maximal accuracy for each complexity. When using diatonic scale degrees as the surface tokens, the functional harmony theory that uses Tonic, Dominant, Subdominant clustering was plotted on the optimal curve, while the Major, Minor, Diminished clustering was far less accurate. This means that the functional harmony theory is more favorable than Major, Minor, Diminished clustering when using the diatonic scale degrees as the surface tokens. In addition, they employed the analysis with automatically labelled data. They adopted the key-detection algorithm of White and Quinn [13] that used the Krumhansl-Shmuckler algorithm [5] on windows of eight slices, and picked up the most common 22 pitch classes (with the bass notes) as the surface tokens. They reported that the obtained clusters were quite close to the Tonic, Dominant, Sub-dominant classification when the number of the categories was 3.

On the other hand, Tsushima et al. [12] found the chord functions in datasets of popular music pieces, using generative models rather than clustering: HMM and Probabilistic Context Free Grammar (PCFG). They reported that when the number of states was 4, the output probability of HMM trained with a popular music dataset could be interpreted as the chord functions: Tonic, Dominant, Sub-dominant, and Others [12], though the model achieved less perplexity with more states. Although PCFG is more advantageous since it can represent more external structures such as long-range dependency of cadence, the reported performance did not exceed that of the HMM. Using a trained HMM as the initial value of PCFG was also found to be clearly effective. However, for the melody harmonization task, PCFG was reported more effective than HMM [11]. For training the HMM, they tested the expectation-maximization (EM) algorithm and Gibbs Sampling (GS) since GS showed significantly higher accuracy than the EM algorithm in the part-of-speech tagging task [2]. They reported that the GS algorithm may perform better especially for a large number of hidden states since it can avoid being trapped in bad local optima.

3 Chord function identification with HMM

Following the previous works, we used a statistical approach to identify chord functions. We chose the HMM for our model because its structure agrees well with that of the functional harmony theory. We expect that the states of the HMM represent chord functions, instead of another possible approach that assumes chord symbols as the hidden states and surface notes as the output tokens.

We obtained the chord functions with the plausible number of states that was fed back by the modulation detection in the following steps.

1. Train the HMM in the range of 2–12 states and choose the best parameterization for each number of states in terms of perplexity.
2. Calculate the likelihood of the chord progression of every candidate partition of key blocks by using the obtained HMM, and determine the intervals of

modulation that maximize the sum of the likelihoods of key blocks by using the set partitioning model.[3]
3. Obtain the best number of states that scores the highest sum of likelihoods on an entire corpus.

3.1 Dataset

We used J. S. Bach's four-part choral pieces BWV253-438 from the Music21 Corpus [1] as our dataset. Several pieces in the chorales have complicated modulations which are not compatible with the modern tonalities. We should also consider that the key signatures of several pieces are different from the modern tonal system. We excluded 24 pieces which obviously differed from the major and the minor key: 22 dorian, 1 mixolydian, and 1 phrygian, and targeted the remaining 94 major pieces and 68 minor pieces. However, there were still pieces that retained the feature of the church modes, especially in minor mode pieces.

To learn the chord functions, we used only the first and the last phrases[4] that were identified by the fermata[5] notation in each piece because we supposed to be able to identify the key of these phrases from the key signature. Those pieces whose first and last chords were different from the tonic of the key signature were excluded.

3.2 Functional chord progression model based on HMM

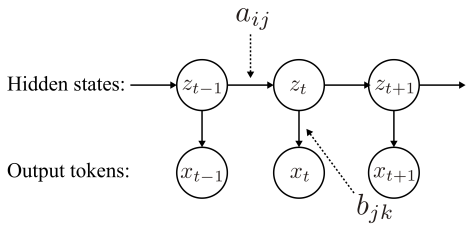


Fig. 1. Graphical representation of the hidden Markov model (HMM).

Model We regarded chord degrees as output tokens for the HMM in Fig. 1, and states as chord functions. Here, z_t denotes the hidden state and x_t the output token at each time step. The state-transition probability is denoted by a_{ij} and the output probability b_{jk}. The number of distinct states is denoted by N_s, and that of output tokens N_v. When we need to specify a state, we use $(z_t =)s_i, i \in \{1, \ldots, N_s\}$, and for output tokens we use $(x_t =)v_k, k \in \{1, \ldots, N_v\}$.

[3] The set partitioning model is a sort of the linear programing.
[4] In this paper, a phrase means a section divided by fermatas.
[5] Fermata is a notation which usually represents a grand pause. However, in the chorale pieces, it represents the end of a lyric paragraph.

Surface tokens We modelled the chord functions of the major key and the minor key by the HMM, and investigated the number of states in the range from 2 to 12. To train the models, we transposed all the major keys to C major and all the minor keys to A minor.

Basically, we used chord degrees on the diatonic-scale as the surface tokens because we trained the models only for C major and A minor, and used them for other keys by transposing the surface tokens. We needed to use more tokens for the minor key considering the all possible chords that were created by introducing the leading-tone in addition to the natural **VII**. The surface tokens of the major and minor keys are listed in Table 1.

Major		Minor	
Chord name	Proportion	Chord name	Proportion
C major(**I**)	30.50%	A minor(**i**)	28.59%
G major(**V**)	19.56%	E major(**V**)	14.94%
D minor(**ii**)	12.35%	C major(**III**)	7.91%
A minor(**vi**)	10.76%	B diminished(**ii**°)	7.22%
F major(**IV**)	9.57%	D minor(**IV**)	6.23%
B diminished(**vii**°)	5.44%	G major(**VII**)	6.13%
E minor(**iii**)	4.37%	G♯ diminished(**vii**°)	5.24%
Others	7.45%	F major(**VI**)	5.14%
		E minor(**v**)	3.46%
		C augmented($\mathbf{III}^{+}$)	2.08%
		Others	13.06%

Table 1. Surface tokens.

Here, we simply removed chords that were not classified to major, minor, diminished, and augmented by using a function to classify the qualities of chords in the Music21 library [1]. We treated the remaining chords that were not in the diatonic scale as 'Others'. In addition, we treated a succession of the same chord as a single surface token.

Optimization method We used the simple EM-based approach known as the Baum-Welch algorithm for learning the HMM parameters from data. While the GS would be effective to avoid bad local optima [12, 2], we rather employed the optimization from a large number of initial values to study the variance of locally optimal parameterizations. For each number of states, we used 1000 different initial values to learn the parameters. We randomly initialized the state-transition probability matrix, while the output probability matrix was initialized uniformly. For each initial value setup, the training data consisting of randomly connected pieces, where we shuffled the opus numbers of the pieces and put them into one sequence.

Evaluation measures We evaluated the parameterizations of the HMM obtained from 1000 different initial values on each number of states (among 2 – 12) to find the optimal one. For each number of states, we selected the optimal parameterization which scored the lowest perplexity defined by following equation:

$$\mathcal{P} = \exp\left(-\frac{1}{|\boldsymbol{x}|}\ln P(\boldsymbol{x}|\boldsymbol{\theta})\right). \tag{1}$$

We also calculated the variance of the 1000 optimal parameterizations for each number of states by employing the K-means clustering around the best optimal parameterization. A large variance indicates larger difficulty to consistently obtain the optimal parameterization.

3.3 Modulation detection as the set partitioning problem

The remaining problem is to select the best number of hidden states. We obtain it by using the modulation detection described below. We select a key that maximizes the likelihood, calculated by the obtained HMM. If we simply apply the HMM, we can only obtain one optimal key for a target piece. By the set partitioning algorithm to detect modulations, we can assign the optimal key blocks to the target piece. The chord functions are expected to work well for detecting a key, especially when there are modulations in the target pieces.

This idea can be formulated as a special case of the set partitioning model, regarding that a music piece is composed of locally optimal key blocks. Here, we use the following notation.

$T = \{1, \cdots, N_t\}$	Serial number of chords in a target sequence
$C = \{1, \cdots, N_c\}$	Set of indices of candidate blocks
$j \in C$	Index of blocks
C_j	Set of chords in candidate block j
e_{ij}	$e_{ij} = 1$ if chord $i \in C_j$ and otherwise $e_{ij} = 0$
d_j $(j \in C)$	$d_j = 1$ if C_j is chosen in the partition and otherwise $d_j = 0$
r_j	Score (the likelihood and penalty) of candidate block C_j

Table 2. Notation in the set partitioning model.

The objective of this set partitioning model is to maximize $\sum_{j=1}^{N_c} r_j d_j$, which means that we select the set of blocks that gives the highest score. The imposed constraints are $\sum_{j=1}^{N_c} e_{ij} x_j = 1, i \in T, d_j \in \{0, 1\}, j \in C$, which means that a surface token must be included in one and only one block.

Since we used only the chords on the diatonic scales, there were many tokens that were classified as 'Others' described in Table 1 when considering all the candidate keys. We imposed penalty on 'Others' tokens. The penalty value was empirically set to $\log(0.01)$.

4 Experimental results

4.1 Evaluation for each number of hidden states

Perplexity For each number of states, we assumed that a parameterization with a lower perplexity is better. With this criteria, we sorted the results by the perplexity and selected the best one in all the results from 1000 initial values. The best perplexity decreased as the number of states increased (Fig. 2). This result is consistent with the previous work that used a popular music dataset [12].

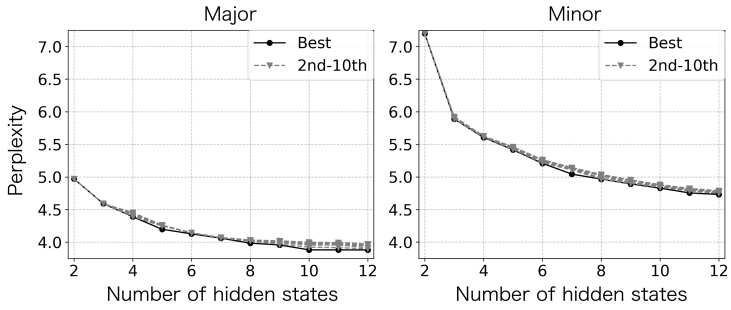


Fig. 2. Perplexities of 10-top parameterizations for each number of states.

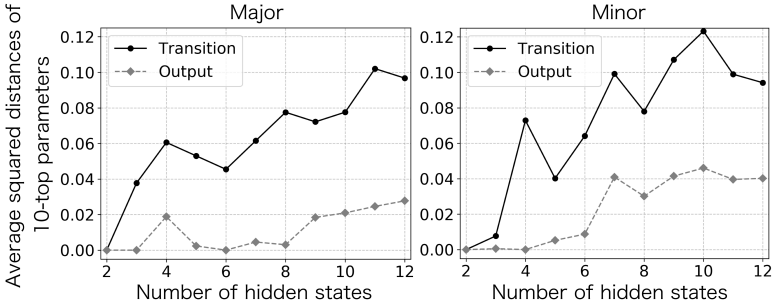


Fig. 3. Average squared distances of 10-top parameters with K-means clustering.

Variance Next, we studied the variance of the optimal parameterizations. For each number of states, we calculated the average squared distances of each of the output and transition probabilities among the top 10 optimal parameterizations. To eliminate the influence of the permutation ambiguity of the state labels, we

adopted the K-means clustering method for calculating the squared distance between two parameterizations of output/transition probabilities. More specifically, we used the Scikit-learn library [7] and fixed the centroids of the clusters as the best optimal parameter values.

As shown in Fig. 3, the distances of the optimal parameterizations increase along with the number of hidden states. This suggests that when the number of hidden states is large there are many different optimal parameterizations and it is difficult to uniquely find the best parameterization solely based on the perplexity.

4.2 Selecting the number of hidden states

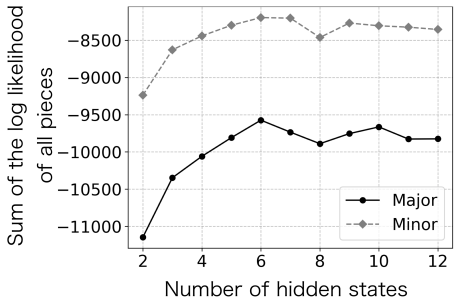


Fig. 4. Sum of the log likelihood of all pieces.

As explained in section 3.3, we obtained the the appropriate number of states by simultaneously employing the chord function identification and modulation detection. To reduce the computation time, we separated a piece into phrases by using the fermata notation, and calculated the likelihood on each phrase.

The 6-state model scored the highest sum of likelihoods both for the major keys and for minor keys (Fig. 4).

4.3 Chord function identification

Major key For the major key, the chords were classified into fine-grained functions, up to 6 states, as shown in Fig. 5. When the number of states is 3, in addition to the clear functions of Tonic $\{\mathbf{I}\}$ and Dominant $\{\mathbf{V}, \mathbf{vii}^{\circ}\}$, there is a mixed function of Tonic and Sub-dominant to which $\{\mathbf{ii}, \mathbf{iii}, \mathbf{IV}, \mathbf{vi}\}$ are assigned. This mixed function is separated into Tonic $\{\mathbf{iii}, \mathbf{vi}\}$ and Sub-dominant $\{\mathbf{ii}, \mathbf{IV}\}$ when the number of states is 4. And then, the state of Dominant is separated into $\{\mathbf{V}\}$ and $\{\mathbf{vii}^{\circ}\}$ with 5 states. Finally, when the number of states is 6, most chords are assigned to an almost unique state, except that $\{\mathbf{iii}, \mathbf{vi}\}$ form one state. Here, we see that $\{\mathbf{iii}\}$ is mainly assigned to Tonic, which recovers the result of Tsushima et al. for popular music datasets [12].

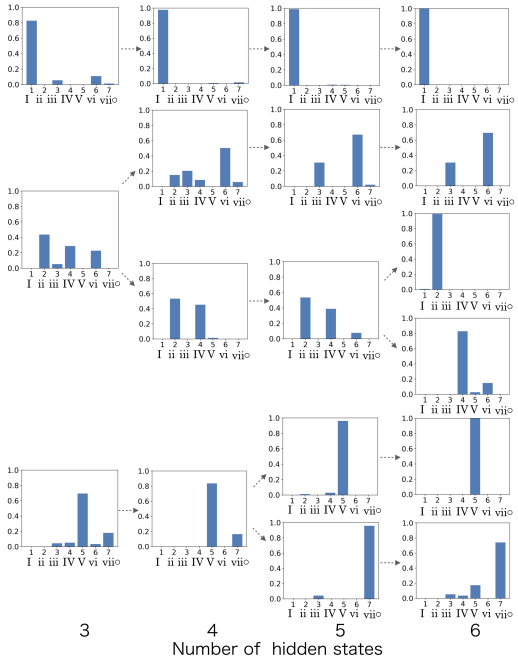


Fig. 5. Output probabilities of the best HMMs for the major key.

The fine-grained state-transition probability is also meaningful. As shown in Fig. 6, we can find detailed functions. For example,

1. The state s_2 for **V** and state s_6 for **vii°** both tend to proceed to state s_4 for **I**, while state s_6 less often proceeds to s_3 for $\{\mathbf{iii}, \mathbf{vi}\}$.
2. Although both states s_1 and s_5 have the function Sub-dominant, s_1 for **ii** more often proceeds to Dominant chords (state s_2 and state s_6) than state s_5 for **IV**.

Minor key The results for the minor key were significantly different from those for the major key, where states corresponding to Tonic and Dominant of the relative major key were obtained when the number of states was larger than 4. With 6 hidden states, in addition to Tonic, Dominant and Sub-dominant, the Tonic of the relative major and that of the Dominant of the relative major were obtained. This result reflects the feature of the choral, whose melodies were composed in the medieval ages in the church modes instead of modern tonalities, prior to the harmonization by J. S. Bach, because the relative keys share the common pitch classes like the church modes.

Rohrmeier et al. also pointed out that the groups of chords corresponding to the relative major key were existing in the minor key clusters [8]. In addition

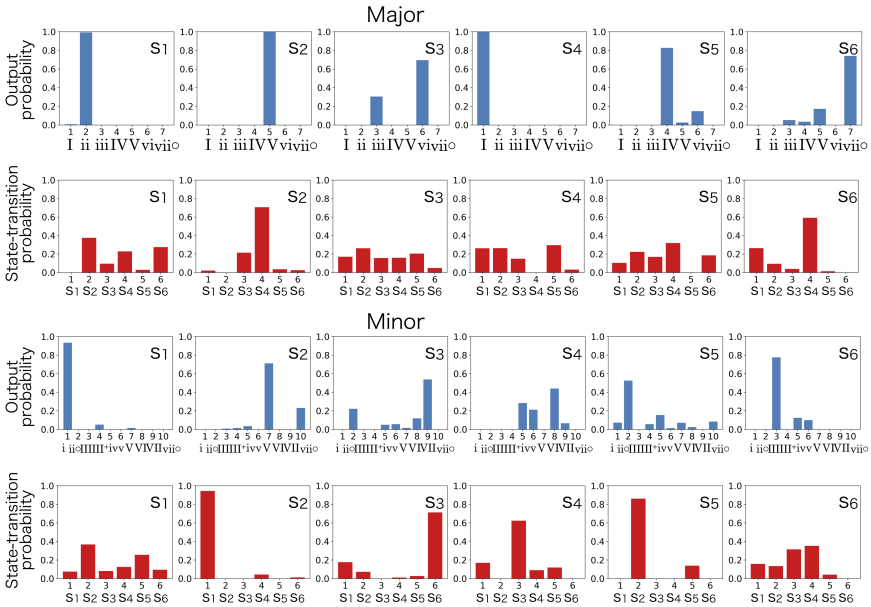


Fig. 6. Output and transition probabilities of 6 hidden states.

to this finding, we found how the same chord could have different functions by observing the value of state-transition probability. As shown in Fig. 6, **ii**° appears in both hidden states s_3 and s_5. Here, state s_5 is Sub-dominant since it tends to proceed to state s_2 which is clearly Dominant. On the other hand, **ii**° in state s_3 can be interpreted as the Dominant of the relative major key since it mainly proceeds to state s_6, which represents **III** corresponding to **I** of relative major key.

4.4 Example of the modulation detection

Although we calculated the sum of the likelihood on separated phrases to reduce the computation time as mentioned in section 4.2, we can detect the modulation on the entire piece. Since pieces of classical music often have a number of modulations and their phrase boundaries are usually not explicitly indicated, this fully dynamic modulation detection is practically useful.

For example, Fig. 7 shows the modulation detection for the piece BWV271. The initial key of this piece is D major, while the key at the end is B minor with a half cadence. This piece has key blocks in D major, B minor, E minor, and A major. The proposed method captured the modulations for the most part.

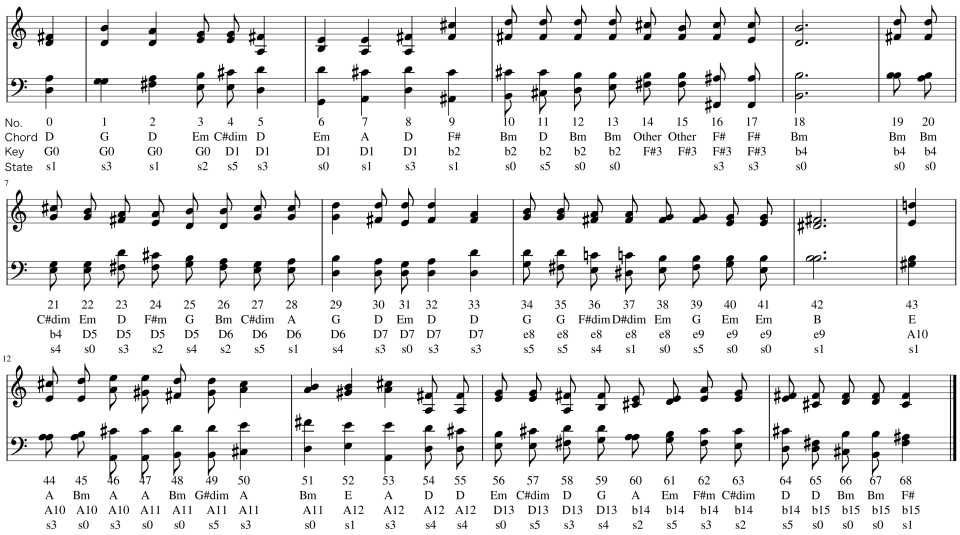


Fig. 7. Modulation detection for the piece BWV271. The 'No.' denotes serial numbers, 'Chord' denotes chord names, 'Key' denotes keys and block numbers obtained by the proposed method, and 'State' denotes HMM state labels.

5 Conclusion

We have employed the Hidden Markov Model (HMM) to identify the chord functions, regarding the surface chord degrees as observed outputs. First, we have looked for the best parameterization for each number of hidden states by perplexity, and then, we evaluated the best likelihood of partitioning by modulation. We found that the most adequate number of hidden states was six, which is not large, and thus we could give the fine-grained interpretations for chord functions; *e.g.*, the Dominant **V** and **vii**$^\circ$ had different tendency towards $\{\mathbf{iii}, \mathbf{vi}\}$, or the subdominant **IV** and **ii** behaved differently toward the Dominant.

We have applied those chord functions to the partitioning by modulation. The interval of modulation was determined dynamically without fixing the scope beforehand, however, the resultant score of partitioning was also fed back to the number of hidden states. Thus, this process is a tandem model, which is one of the most important features of our work.

Another important feature is the characterization of music styles by parameters. In our example, the set of parameters reflects the specific feature of Bach's chorales, where the basic melodies are of church modes while the harmonization is in the Baroque style. In general, other sets of parameters may have a potential to characterize different music styles such as post-romanticism.

Since our main objective was the key identification, we excluded those borrowed chords and assigned an artificial penalty value to them. Thus, to investigate the key recognition with extraneous chords is our immediate future work. And also, the evaluation with human annotations is our another important future

work, even though the human recognition of modulations could admit multiple interpretations. In addition, although we have realized an efficient modulation detection, our method included such errors to regard groups of chords as modulation. To solve this issue, we plan to introduce the notion of dependency in chords, that is to assess the prolongation of the influence of preceding chords.

References

1. Cuthbert, M. S., Ariza, C.: music21: A toolkit for computer-aided musicology and symbolic music data. In: 11th International Society for Music Information Retrieval Conference, pp. 637–642 (2010)
2. Goldwater, S., Griffiths, T. L.: A fully Bayesian approach to unsupervised part-of-speech tagging. In: 45th Annual Meeting of the Association of Computational Linguistics, pp. 744–751 (2007)
3. Hu, D. J., Saul, L. K.: A Probabilistic Topic Model for Unsupervised Learning of Musical Key-Profiles. In: 10th International Society for Music Information Retrieval Conference, pp. 441–446 (2009)
4. Jacoby, N., Tishby, N., Tymoczko, D.: An Information Theoretic Approach to Chord Categorization and Functional Harmony. J. New Music Res., vol. 44(3), pp.219–244 (2015)
5. Krumhansl, Carol L. and Kessler, E. J.: Tracing the dynamic changes in perceived tonal organisation in a spatial representation of musical keys Key-Finding with Interval Profiles. Psychological Review, vol. 89(2), pp.334–368 (1982)
6. Lerdahl, F.: Tonal pitch space. Oxford University Press (2004)
7. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. J. Machine Learning Res., vol. 12, pp.2825–2830 (2011)
8. Rohrmeier, M., Cross, I.: Statistical Properties of Tonal Harmony in Bach's Chorales. In: 10th International Conference on Music Perception and Cognition, pp. 619–627 (2008)
9. Sakamoto, S., Arn, S., Matsubara, M., Tojo, S.: Harmonic analysis based on Tonal Pitch Space. In: 8th International Conference on Knowledge and Systems Engineering, pp. 230–233 (2016)
10. Temperley, D.: The Tonal Properties of Pitch-Class Sets: Tonal Implication, Tonal Ambiguity, and Tonalness. Computing in Musicology, vol. 15, pp.24–38. Center for Computer Assisted Research in the Humanities at Stanford University (2007)
11. Tsushima, H., Nakamura, E., Itoyama, K., Yoshii, K.: Function- and Rhythm-Aware Melody Harmonization Based on Tree-Structured Parsing and Split-Merge Sampling of Chord Sequences. In: 18th International Society for Music Information Retrieval Conference, pp. 502–508 (2017).
12. Tsushima, H., Nakamura, E., Itoyama, K., Yoshii, K.: Generative statistical models with self-emergent grammar of chord sequences. J. New Music Res., vol. 47(3), pp.226–248 (2018)
13. White, C. W., Quinn, I.: The Yale-Classical Archives Corpus. Empirical Musicology Review, vol. 11(1), pp.50-58 (2016)

(Re)purposing Creative Commons Audio for Soundscape Composition using Playsound

Alessia Milo[1], Ariane Stolfi[2], and Mathieu Barthet[1][0000−0002−9869−1668]

[1] Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
miloalessia@gmail.com; m.barthet@qmul.ac.uk
[2] Federal University of Bahia, Salvador, Brazil
arianestolfi@gmail.com

Abstract. Playsound is an open-source web-based interface allowing users to search for, edit and process Creative Commons (CC) sounds from Freesound. In this paper, we present the results from a user study conducted with 17 music production students who created short soundscape compositions only using CC-licensed audio retrieved with tools including Playsound. The students completed an online survey which included the System Usability Scale (SUS) and Creativity Support Index (CSI) questionnaires and open-ended questions. Although Playsound was found helpful to predict how various sounds would blend together and sketch musical ideas, the results suggest that usability and specific creativity factors (exploration and expressiveness) should be improved. We discuss Playsound's strengths and weaknesses and provide insights for the design of tools to support soundscape composition using crowdsourced audio.

Keywords: Creative Commons · Soundscape composition · Freesound · User evaluation, · Creativity support · Web Audio

1 Introduction

Soundscape compositions are musical compositions seeking to elicit listeners' reflections on the interrelationships between sound, nature and human society [15, 21]. It is difficult to define soundscape composition unequivocally. The term *soundscape* is ambiguous probably due to its re-appropriation in several domains to denote as varied artefacts as field recordings, musical compositions, mobile soundtrack, sound designed pieces for theatre, games, movies (etc.) [12]. Acoustic ecology is concerned with the study of *soundscapes* defined as the *"acoustic environment as perceived or experienced and/or understood by a person or people, in context"*[3]. *Soundscape compositions* create the sensation of experiencing a particular acoustic environment through the use of found sounds[4] which are

[3] https://www.iso.org/obp/ui/#iso:std:iso:12913:-1:ed-1:v1:en
[4] sounds of natural objects belonging to the environment rather than from crafted musical instruments

edited and processed by composers in the studio to convey specific artistic meaning. Composers have developed different soundscape composition strategies and styles over time [22, 23], employing a range of aesthetic forms from figurative connotations of the acoustical environment using transparent editing to more abstract references involving advanced audio processing [13]. Audio recordings are a cornerstone of the soundscape composition process during which, as advanced by Westerkamp, *"the artist seeks to discover the sonic/musical essence contained within the recordings and thus within the place and time where it was recorded. The artist works with the understanding that aesthetic values will emerge from the recorded soundscape or from some of its elements"* [23]. For example, Luc Ferrari's soundscape composition *Presque rien, numéro 1* is centred around discernible field recordings of a fishing village during which rhythms emerge from objects or animals such as boat engines and cicadas. With the granulation technique proposed by Truax, which consists in stretching a sound using variable-rate time shifting, audio samples as short as 150ms can yield very rich abstract textures, as exemplified in his piece *The Wing of Nike* which used brief male and female phonemes as source material [20].

With the advent of web technologies and social media, the twentieth century has seen a burst in the sharing culture of digital media (blogs, photos, videos, sounds, etc.). Audio platforms such as Freesound [1, 10] have emerged whose content is populated by a wide range of users, from amateurs to professionals, with self-produced recordings (*crowdsourced* audio). Although this is a rather recent phenomena in the history of music, millions of audio recordings can already be accessed openly through the web (outside of commercial streaming music platforms), spanning a wide range of material from raw field recordings to composed musical samples and sound designed audio effects, to name a few. A large part of this content is shared under Creative Commons (CC) license[5] which provides a legal framework enabling authors to choose reusability conditions, alleviating issues around sealed copyrighted content. However, several barriers remain to facilitate the reuse of CC audio content for media production: (i) web-based CC audio platforms are disparate and there is a lack of unified search mechanism for users; (ii) the metadata associated to CC audio content suffers from issues of sparsity and noise due to the lack of editorial curation (crowdsourced metadata), yielding a certain amount of incomplete or erroneous bibliographic information and recording descriptions, (iii) there is a lack of tools supporting sound designers or composers to access and integrate CC audio content into their productions. The recent Audio Commons initiative[6] has instantiated an ecosystem attempting to solve some of these issues using semantic audio and tools supporting creativity [24]. In this work, we investigate how crowdsourced CC audio content may benefit the task of soundscape composition. We use Playsound, an open access browser-based tool [18] which enables users to query and play sounds from Freesound. In the remainder, after discussing related

[5] https://creativecommons.org/
[6] https://audiocommons.org/

works, we detail a user study conducted to assess how Playsound supported the search and/or generation of sounds for soundscape composition.

2 Related work

Soundscape composition has been investigated in the musical metacreation field concerned with the computational generation of music [8]. Generative music systems can be distinguished based on the degree of autonomy they provide to the user/composer, from fully algorithmic systems where control over the outcome is minimal and left to a few parameters to interactive composition systems providing more agency to composers. [19] developed an automatic sound recommendation technique for soundscape composers given a seed textual description, by using natural language processing and query expansion with Twitter. [8] devised a real-time composition system in which agents interact in musically intelligent ways according to psychoacoustic descriptors and criteria informed by human soundscape composition strategies. Other works have proposed a data-driven approach (so-called "datascaping") by merging automatic sonifications of environmental data with found sounds characteristic of the place and activities being represented [14]. Soundscapes have also been the object of research in networked music performance with studies and performances investigating how geographically displaced creators can collaborate to generate shared soundscapes over Internet [2, 5].

The present study focuses on soundscape compositions as linear and fixed media devised in the studio by individual composers. We do not attempt to automate certain processes of musical creativity or support remote musical interactions, rather we investigate whether crowdsourced audio and tools to access it can help composers producing soundscapes with standard digital audio workstations (DAWs). We focus on a specific tool which we devised in previous work, Playsound.space[7] [16], a web interface providing a fast access to the Freesound audio database [1, 10] using Web Audio, a high-level JavaScript API for processing and synthesizing audio in web applications[8]. Playsound has been used for live music improvisation [17] and interactive composition [18], while we consider it here in a non live scenario, to support the creative process of soundscape composition. Other tools leveraging Freesound have been proposed such as the Freesound Explorer, a visual interface providing a two-dimensional space for content exploration and music making by linking content in the space [9].

3 Methods

3.1 Playsound

Playsound was initially designed as a music making tool enabling non musicians and musicians to perform music using ubiquitous technologies (e.g. laptop with Internet) and pre-existing content (crowdsourced audio) [16]. It serves two main functionalities, to search for Creative Commons sounds pooled from Freesound,

[7] http://playsound.space
[8] https://www.w3.org/TR/webaudio/

Fig. 1. Playsound interface: it includes a search bar (top), a sound selection grid showing spectrograms (middle), and sound players to create mixes (left).

and to produce new material by mixing sounds together. The user interface (UI) of Playsound is shown in Fig. 1. Users can enter semantic queries (e.g. "soundscape") using the text bar on the top. Queries can be translated in different languages (English is predominantly used on Freesound). The sound results are displayed in a grid fashion with a maximum of 40 sounds per page and users can browse across pages using the arrows at the top. Each sound is represented by its spectrogram (time-frequency representation) retrieved from Freesound, as well as its filename. Sounds can be audited individually or conjunctly by pressing the play icons appearing at the top of the spectrograms. The panel to the left of the interface provides a stack of audio players each having mixing and processing controls, as shown in Fig. 2. These controls enable users to select for each selected sound, a start and end point, to activate or deactivate looping, and adjust the volume, playback speed and panning. Playsound currently provides access to content under CC0 (no rights reserved) and CC-BY licenses[9] which let others distribute, remix, tweak, and build upon a work, even commercially, as long as credits are given to the original creator. Sounds can be downloaded by clicking on the top right corner icon in the audio player which redirect users to the original sound page on Freesound (Freesound requires authentication to download sounds[10]). Playsound also enables to record the mix produced with the tool. Authorship information for CC-BY sounds which require credits are displayed at the bottom of the UI.

[9] https://creativecommons.org/licenses/
[10] https://freesound.org/home/register/

In this study, we do not assess Playsound as a soundscape composition tool but rather as a tool to support the soundscape composition process by providing access to audio material that can be further arranged and processed in the DAW.

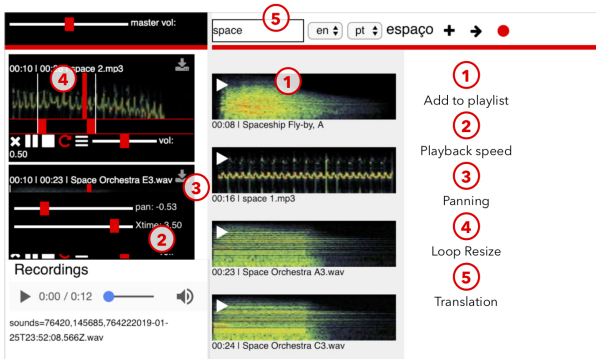


Fig. 2. Playsound audio editing and processing features

Playsound relies on a client/server architecture. The client works on laptop or desktop computers with most of the modern browsers compatible with Web Audio[11]. The Playsound server processes user queries and is in charge of the authentication to Freesound. Communication with Freesound is made through a RESTful (representational state transfer) API (application program interface). Playsound has been released as an open source software[12].

3.2 Context and procedure

Fig. 3. Photos of soundscape training sessions at QMUL

[11] Although Playsound can potentially work on mobile devices, it is, at the time of writing, not optimised for this type of platform

[12] https://github.com/arianestolfi/audioquery-server

The study was conducted with students of the Sound Recording and Production Techniques (SRPT) module from Queen Mary University of London during the autumn of 2018. During the module, the students were first introduced to the physics of sound, acoustic ecology, psychoacoustics and listening through excercises. Amongst the pedagogical activities, the students participated in soundwalks [7] during which they were invited to describe the sounds they perceived while blindfolded, while others took notes on their auditory observations (see Fig. 3). The students followed practical sessions on microphone techniques in the studio and outdoor. They received training on Apple's Logic Pro X digital audio workstation and mixing techniques. After lectures on soundscapes and compositional strategies, students had three weeks to produce a short soundscape on a specific theme. This soundscape composition task acted as a formative assessment focusing on the practice of audio editing, processing and mixing. To this end, students had to use prerecorded audio licensed under Creative Commons. They were introduced to Playsound as well as other tools from the Audio Commons Initiative including Jamendo's Sound and Music Search Tool[13], Le Sound's AudioTexture plugin[14] and Waves Audio's SampleSurfer plugin [24]. Documentation and tutorial videos were provided[15] and students could ask questions during the module's help sessions. The soundscape themes were generated in class following a process inspired by the structured brainstorming technique *bootlegging* [11], which involved to randomly combine ideas to form unexpected juxtapositions which became the basis of a concept. Students had to write down on post-its one idea in each of four categories: *character*, *place/environment*, *situation/action*, and *mood*. Post-its were shuffled and students had to randomly pick one up per category. After combining the ideas, a brainstorming session was conducted to start forging creative narrative for the soundscape. Students could then refine their concepts in their own time while working with the recordings. Along with their compositions, students had to submit a report which described their creative and technical decisions and discussed the technologies and content used during production. After the submission, students were invited to take part in the survey described in Section 3.3 which was non compulsory and anonymous. A follow-up "soundscape walkthrough" session was organised during which students presented and played their soundscapes.

3.3 Online survey

An online survey[16] was designed to collect feedback from the participants after the completion of the soundscape composition task. The survey included demographics questions and separate sections to assess each of the tool (we focus in this paper on the results obtained for Playsound). Tools were assessed using a

[13] https://audiocommons.jamendo.com/
[14] https://lesound.io/product/audiotexture/
[15] Playsound demo video: https://tinyurl.com/playsounddemo
[16] The survey was implemented using JISC's online survey tool: https://www.onlinesurveys.ac.uk/

combination of well-established human-computer interaction (HCI) metrics and questions specifically designed for the study. The HCI metrics comprised the System Usability Scale (SUS) [4] which consists of a set of ten 5-point Likert items, and the Creativity Support Index (CSI) [6], which includes twelve 11-point Likert items. The SUS provides a quick assessment of the usability of a system on a scale from 0 to 100. This framework has been used in over 500 studies[17] which yielded an average SUS of about 68, which can serve as overall benchmark. The CSI assesses six orthogonal factors, each of them being assigned a score: *exploration* (to what extent it is easy to explore ideas, options, designs or outcomes), *expressiveness* (being able to be expressive and creative while doing the activity), *enjoyment* (the level of enjoyment or engagement of the activity; the CSI authors decided to use 'enjoyment' and 'engagement' indistinctly because they were not found to be orthogonal), *immersion* (the extent to which the tool is transparent and the attention is focused on the activity more than on the system or tool used), *collaboration* (whether the system allows to collaborate between people), and *results worth effort* (the relation between the effort and the final product outcome is worth the effort). The CSI questionnaire also includes pairwise comparisons assessing the relative importance of each of the six creativity factors for the task (factor weights). The CSI survey metric is computed by taking into account both the factor scores and their importance and ranges between 0 and 100. The next part of the survey consisted in task-specific Likert items (L) and open-ended questions (Q). Likert items assessed (L1) users' interest in the tools, (L2) integration in the workflow, (L3) support for CC license information, (L4) relative use compared to other tools, (L5) ability to find targeted sounds or discover unforeseen ones, (L6) usage of audio editing and (L7) processing, and (L8) ability to create new relevant audio material. Open-ended questions aimed at better understanding (Q1) why and how the tool was used, (Q2 and Q3) what were the perceived advantages and disadvantages of the tools, (Q4) which improvements they envisioned, (Q5) what types of sounds they used, and for Playsound, (Q6) how the spectrogram helped them choosing sounds. Finally, they could provide overall comments (Q7). Survey sections were randomised across participants to avoid potential order effects. The survey lasted about 1h.

4 Results

4.1 Participants

The survey was completed by 17 participants (8 females, 9 males) from the SRPT module described in Section 3.2. The average age was 25.5 years (SD = 4.7). According to self-reports on music production expertise, the group was composed of 10 novices, 5 intermediates, and 2 expert users.

4.2 Statistical analyses

Statistical analyses were conducted using a Type I error of 5%. The following notations are used: M (mean), SD (standard deviation), N (number of participants), and S.E. (standard error around the mean). The task-specific Likert items

[17] https://measuringu.com/sus/

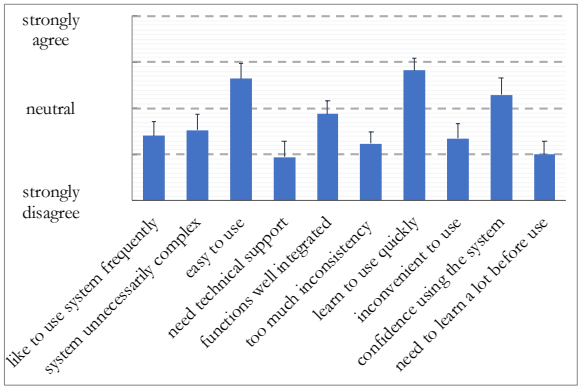


Fig. 4. Results for SUS items.

were subjected to Mann-Whitney-Wilcoxon (MWW) tests to assess the effects of gender, age, and music production experience. Only one significant effect of experience was found (p=0.048) for the statement *"I have created new sounds or new sequences using the functions of Playsound"* to which novices agreed to a larger extent (M=4.67, SD=2.55) than intermediate/experts (M=2, SD=2.88). This suggests that as a tool to generate new material, Playsound appears more suitable for non specialists, which is line with the initial design intention [16].

4.3 System Usability Scale and Creativity Support Index

Playsound obtained a mean SUS score of 62.5/100 (SD=26.7) which is slightly below the average discussed in Section 3.3. Fig. 4 shows how the participants rated Playsound on the SUS items. Participants slightly agreed that Playsound was quick to learn and easy to use. They also tended to disagree that it needed further technical support or extensive training. Participants tended not to find the tool inconvenient and inconsistent, and also unnecessarily complex. However, there was only a slight agreement about the confidence in using the system, and also a slight disagreement in wanting to use Playsound frequently. No clear opinion was forged on whether the functions were well integrated.

The CSI was computed using Python open-source tools[18]. Playsound obtained a mean CSI of 46.6 which indicates some shortfalls. The analysis of the six CSI creative factors indicates that Exploration (M=4.12/5) and Expressiveness (M=3.88/5) were judged the most relevant factors for the soundscape composition task. The tool obtained highest scores for Results Worth the Effort (M=5.29/10) and Exploration (M=4.85/10). The three highest overall weighted factor scores were Exploration (M=19.3/50), Expressiveness (M=17.9/50), and Results Worth the Effort (M=14.1/50). The other factors were ordered as follows: Immersion (M=8.9/50), Enjoyment (M=8.4/50), and Collaboration (M=1.4/50).

[18] https://github.com/axambo/hci-python-utils

Since the soundscape composition task was individual, the low score for Collaboration is legitimate and its influence on the CSI is limited by a low factor weight.

4.4 Thematic Analysis

We conducted a thematic analysis [3] of the answers to the 8 open-ended questions described in Section 3.3 using an inductive approach. We first familiarised with the data and then assigned semantic codes to each answers. Themes were further searched, reviewed and finalised. The thematic analysis was performed by two coders (the first and last authors) and results were integrated after revising codes and themes. We present below a report for each of the 12 themes organised by decreasing order according to code occurrences reported in brackets (305 codes in total).

Spectrogram (27 codes). Spectrogram was the theme which obtained the largest number of codes. The large majority of participants found the spectrogram visualisations provided by Playsound very valuable: it helped them predict how the audio content would sound during selection (11 codes), providing information about the rhythm, dynamics, texture, strength, stability, the frequencies and pitch register. The spectrogram helped to identify, compare sounds and find similar match. It also helped them visualise and predict the amount of noise (3 codes), perform editing (2 codes) by finding the start and end points or sections, and increased efficiency (1 code). For a small amount of participants (4 codes), spectrograms were found not useful or provded too much information. **Experimentation and Sketching** (24 codes). A large number of participants reported finding Playsound useful for experimentation and sketching. The tool was commended for its ability to easily and quickly combine sounds (14 codes) and evaluate them for *"drafting ideas before heading into the DAW"*. This was supported by the possibility of auditioning (6 codes) multiple samples and making comparisons supported by the spectrogram visualisations. The recording function (4 codes) was found *"convenient"* to capture a mix or individual sounds. Participants also liked to be able to download (3 codes) individual samples or the mix *"in one go"*. **Improvements** (23 codes). Many suggestions for improvements were made. Some concerned the design of the user interface (8 codes), for example to avoid the duplication of audio players for the same sound (however this enables to apply different edits/processing to the same sound), to provide visual feedback when sounds are buffering or being processed, and to improve the clarity of the metadata. Several participants expressed interest for a complementary environment (5 codes) allowing them to *"to develop an idea further when sounds have been decided, but with a similar approach (simple, quick, halfway between current playsound and the DAW)"*, such as an arrangement window. Some participants were keen to have more audio effects (4 codes) such as EQ, compressors and delays, and to import their own tracks or export a mix as multitrack (2 codes). A participant suggested an offline version working with pre-loaded samples. **Search** (18 codes). Both limitations (10 codes) and advantages (8 codes) were reported for the sound search functionality of Playsound. Several participants surprisingly commented on the limited number

of sounds available and wished to access more sounds. This may be partly attributed to the lack of search filters making it complex to find relevant sounds as it requires to *"trawl through search results"* (e.g. short sounds are presented first and longer sounds are often on other pages). Adding search filters (e.g. *"popularity/relevance/duration"*) was deemed important for the soundscape composition task. Several other participants liked the richness of the available content (which could not be found in other tools), the simple and quick search process, and the relevance of the results. **Disinterest** (12 codes). Several participants expressed a disinterest in the tool for the task in comparison to the other tools. This was explained by usability issues (see UI theme) and/or a preference for other tools. **Intuitiveness** (10 codes). In contrast to the previous theme, several participants found the tool to be *"easy"*, *"simple"*, *"straightforward"*, *"intuitive"*, *"effective"*, and *"fun"* (e.g. D24 *"it was a useful way of visualising freesound search results which was more intuitive and fun than usual"*). **User interface** (9 codes). The design of the UI received a number of negative comments, with participants suggesting to improve its appearance, usability, and clarity. One participant thought that the design was too much focused on the spectrogram component and not enough on the mixing aspects (D19: *"I think it could do with better design for the layering/combination/basic effects - there is a weirdly large amount of focus on the spectrogram browser component, as opposed to the sound listing/playback."*). In the same vein, another participant found that the audio processing functions were not obvious. **Bugs** (9 codes). Several participants experienced issues with certain browsers (e.g. Safari) or specific functionalities which afffected usability and creativtiy support. Loading longer samples caused some bugs and removing a sound before it finished caused the sample to run indefinitely in the background. **Audio editing** (7 codes). Two trends appeared when gauging the audio editing capabilities of the tool, one group who liked its simplicity and another who found that it was too basic or difficult to use. **CC license** (7 codes). Some participants found that the CC attribution information was clearly reported but several mentioned that it was difficult to find, or that there was a lack of CC license information. Suggestions for improvements included to export referential information into files. **Openness** (3 codes). The openness of the tool was commended by some participants (D19: *"Playsound acts like a referencing tool, tracking a collection of samples and making it painless to share this list via a basic URL between devices. Without access to a Logic license, being web accessible is very useful for quickly evaluating multiple samples alongside one another (early prototyping)."*). **Latency** (2 codes). A small number of comments reported that occasionnally the tool was slow at loading the desired sounds.

5 Discussion

The quantitative and qualitative analyses indicate a polarisation of the participants, those who engaged positively with Playsound and commended its intuitiveness and ability to quickly audition multiple sounds and sketch ideas, and others who disliked the UI or found the tool too limited compared to other tools for soundsape composition. Two main reasons can be stipulated to explain cur-

rent drawbacks to support soundscape composition: (i) given that participants produced their compositions in the DAW, preference leaned towards tools providing access to CC content from within the DAW, such as the AudioTexture and SampleSurfer plugins, (ii) given the importance of search filters and metadata to select relevant recordings with good audio quality for soundscape composition, participants tended to prefer tools providing a richer amount of information about the sounds (Sound and Music Search Tool). The apparent limitation in number of sounds available expressed by some participants may have been a side effect of UI design issues regarding search and the way to browse the results (icons to browse pages at the top of the UI are mixed with those related to search and recording). Compared to the other tools, one of the strengths of Playsound was its spectrogram representation which was found very useful by a wide majority to predict certain acoustical qualities.

It is interesting to note that in a task of collaborative free music improvisation [16], Playsound obtained both a high SUS score (M=82.5/100, SD=8.94) and CSI (M=71.7, SD=15.6). For collaborative free music improvisation, Expressiveness and Exploration were also the two factors judged most important by participants, followed by Immersion and Collaboration. The quick responsiveness of Playsound and simple audio players suits well live music application using non metric structures. The strengths of Playsound for liveness and simple editing and processing functionalities may become weaknesses in the crafting of soundscape compositions spread over several weeks; some participants wished to be able to develop ideas further, and for this, the models adopted by DAWs are probably more appropriate. Functions to edit the sound content and vary the playback speed have been introduced, reducing also browser compatibility. However, such audio manipulations do not seem sufficient for the soundscape composition practitioner, as illustrated by the fact that some participants would have liked to be able to organise the sounds in an arrangement window and synchronize them more precisely. Apart from the implementation of an interactive timeline, which would require a completely new design, a possible solution could be to add more controls to position loops in time. In order to overcome some of the technical limitations related to playback which could undermine the responsiveness of the interface. We foresee that the integration of modular components, such as adaptive buffering and loading status icons, would improve usability and confidence in using the tool. Interestingly, several participants have found a value in using the Playsound interface to reference authors and easily share Creative Commons content, thanks to its capability of specifying sounds directly within URLs.

6 Conclusions

This study investigated how Playsound - a web interface initially created for live improvisation reusing CC audio content - can support content search and ideation in soundscape composition. Survey participants generally found the tool simple and relevant for auditioning sounds (e.g. to predict how various sounds would blend together through mixing and spectrogram visualisations), but limitations occurred for example due to the impossibility to arrange content in a

highly organized way as in a DAW. Participant feedback helped to identify some directions of development to improve the usability and better support creativity both in performative and studio practice, for the quick sketching and sharing of musical ideas. If the exploration creativity factor was favourably supported on overall, probably due to the richness of CC content, more work is needed to improve the expressiveness of the tool, a factor deemed important for soundscape composition involving the manipulation of prerecorded audio material.

7 Acknowledgments

This work was supported by the EU H2020 grant Audio Commons Initiative (No. 688382) and the Centre for Digital Music at QMUL. We also acknowledge support from University of São Paulo's NuSom Research group and the CAPES PDSE grant awarded to Ariane Stolfi.

References

1. Akkermans, V., Font, F., Funollet, J., de Jong, B., Roma, G., Togias, S., Serra, X.: Freesound 2: An improved platform for sharing audio clips. Proc. ISMIR (2011)
2. Barbosa, Á.: Displaced soundscapes: A survey of network systems for music and sonic art creation. Leonardo Music Journal pp. 53–59 (2003)
3. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2) (jan 2006), http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa
4. Brooke, J., et al.: Sus-a quick and dirty usability scale. Usability evaluation in industry **189**(194), 4–7 (1996)
5. Brown, N., Chudy, M., Alarcon, X., Papadomanolaki, M.: Transglasphone. In: 12th International Symposium of Computer Music Multidisciplinary Research: Bridging People and Sound (Music Program). São Paulo, Brazil (2016)
6. Cherry, E., Latulipe, C.: Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. ACM Transactions on Computer-Human Interaction **21**(4), 1–25 (2014)
7. Drever, J.L.: Soundwalking in the city: a socio-spatio-temporal sound practice. In: 5th Int. Symposium on Temporal Design (2011)
8. Eigenfeldt, A., Pasquier, P.: Negotiated content: Generative soundscape composition by autonomous musical agents in coming together: Freesound. In: ICCC. pp. 27–32 (2011)
9. Font, F., Bandiera, G.: Freesound explorer: Make music while discovering freesound! In: Proc. WAC (2017)
10. Font, F., Roma, G., Serra, X.: Freesound technical demo. Proc. ACM - MM '13 (2013)
11. Holmquist, L.E.: Bootlegging: Multidisciplinary brainstorming with cut-ups. In: Proc. 10th Conf. on Participatory Design 2008. pp. 158–161. Indiana University (2008)
12. Kelman, A.: Rethinking the soundscape: A critical genealogy of a key term in sound studies. The Senses and Society **5**, 212–234 (07 2010). https://doi.org/10.2752/174589210X12668381452845
13. Loufopoulos, A., Mniestris, A.: Soundscape models and compositional strategies in acousmatic music. Soundscape: The Journal of Acoustic Ecology **11**(1), 33–36 (2012)
14. Pigrem, J., Barthet, M.: Datascaping: Data Sonification as a Narrative Device in Soundscape Composition. In: Proc. Audio Mostly (2017)
15. Schafer, R.M.: The soundscape: Our sonic environment and the tuning of the world. Simon and Schuster (1993)
16. Stolfi, A., Ceriani, M., Turchet, L., Barthet, M.: Playsound.space: Inclusive Free Music Improvisations Using Audio Commons. In: Proc. Int. Conf. on NIME (2018)
17. Stolfi, A., Milo, A., Ceriani, M., Barthet, M.: Participatory musical improvisations with playsound. space. In: Proc. WAC (2018)
18. Stolfi, A., Milo, A., Viola, F., Ceriani, M., Barthet, M.: Playsound.space: An ubiquitous system in progress. In: In Proc. of VIII UbiMus Workshop (2018)
19. Thorogood, M., Pasquier, P., Eigenfeldt, A.: Audio metaphor: Audio information retrieval for soundscape composition. In: Proc. SMC. pp. 277–283 (2012)
20. Truax, B.: Discovering inner complexity: Time shifting and transposition with a real-time granulation technique. Computer Music Journal **18**(2), 38–48 (1994)
21. Truax, B.: Acoustic communication, vol. 1. Greenwood Publishing Group (2001)
22. Truax, B.: Genres and techniques of soundscape composition as developed at simon fraser university. Organised sound **7**(1), 5–14 (2002)
23. Westerkamp, H.: Linking soundscape composition and acoustic ecology. Organised Sound **7**(1), 51–56 (2002)
24. Xambó, A., Font, F., Fazekas, G., Barthet, M.: Foundations in Sound Design for Linear Media: An Interdisciplinary Approach, chap. Leveraging Online Audio Commons Content For Media Production. Routledge (2019)

Generating Walking Bass Lines with HMM

Ayumi Shiga[1] and Tetsuro Kitahara[1] *

College of Humanities and Sciences, Nihon University, Japan
{shiga,kitahara}@kthrlab.jp

Abstract. In this paper, we propose a method of generating walking bass lines for jazz with a hidden Markov model (HMM). Although automatic harmonization has been widely and actively studied, automatic generation of walking bass lines has not. With our model, which includes hidden states that represent combinations of pitch classes and metric positions, different distributions of bass notes selected at different metric positions can be learned. The results of objective and subjective evaluations suggest that the model can learn such different tendencies of bass notes at different metric positions and generates musically flowing bass lines that contain passing notes.

Keywords: Jazz bass, automatic generation, hidden Markov model

1 Introduction

Creating walking bass lines is a fundamental skill required for jazz bassists, because bass lines in general are not explicitly described in musical scores that jazz musicians use. Since each score that they use, called a *lead sheet*, includes only a dominant melody and chord progression, they have to create a musically appropriate bass line from that melody as well as the chord progression to play jazz bass. However, that process is challenging for novice bassists, because it requires constructing bass lines that satisfy both simultaneity (i.e., harmonic congruency between the bass line and chord backing) and sequentiality (i.e., smooth succession of notes within the bass line).

Although many systems for harmonization—that is, for outputting chord progressions or four-part harmonies—have been developed [1–6], few have attempted to generate walking bass lines. Dias et al. [7] developed a walking bass line generator following the contour-based approach, in which the user can specify whether the pitch of the bass line ascends or descends according to two parameters: direction and openness. Whereas *direction* specifies whether the first bass note in the next chord is higher or lower than that in the current chord, *openness* specifies how directly the bass line progresses from the first bass note in the current chord to that in the next chord. In particular, *low openness* indicates a direct path—for example, C–D–E–E–F—whereas *high openness* indicates an

* This project was supported by JSPS Kakenhi (JP16K16180, JP16H01744, JP17H00749, and JP19K12288) and the Kawai Foundation for Sound Technology and Music.

indirect path—for example, C–E–G–G$^\flat$–F. When choosing passing notes, the system applies the idea that stronger beats tend to have chord tones and that the last beat of each bar can have a chromatic approximation to the first note in the next bar.

In other studies, Kunimatsu et al. [8] developed a system that automatically composes pieces of blues music consisting of a melody, a chord progression, and a bass line using genetic programming. To generate a bass line, bass line candidates are evaluated in a fitness function based on integrity with the chord progression and music entropy. Ramalho et al. [9] developed a jazz bass player agent that memorizes existing bass-line fragments and reuses them by means of case-based reasoning to generate bass lines. Meanwhile, Piedra [10] developed a bass line generating agent by using a probabilistic model to extract musical knowledge from a collection of MIDI-based bass line loops, although he focused on electronic dance music, and thus did not consider walking bass lines for jazz. Indeed, no researchers have previously attempted to generate walking bass lines for jazz from a data-driven probabilistic model.

In this paper, we propose a method of generating walking bass lines using a hidden Markov model (HMM). HMMs allow estimating the most likely sequence of hidden states from a sequence of observed symbols and are commonly used in *harmonization*—that is, creating a chord progression for a given melody—in which a *chord progression* is a sequence of hidden states, whereas a *melody* is a sequence of observed symbols (e.g., [3]). HMMs are widely considered to be good models for learning relations between a chord progressions and a melody to be performed under the chord progression. Therefore, despite the lack of previous attempts to do so, we expected our HMM to be effective in generating a walking bass line for a given chord progression.

2 Method

2.1 Problem statement

Our aim was to generate a bass line for a given chord progression. Ideally, a melody should also be considered, because the most musically appropriate bass line depends on the melody even if the chord progression is the same. However, we did not consider melodies in our work given the lack of any database of walking bass lines involving dominant melodies. For simplicity's sake, we assumed chord progressions to have only one chord for each bar, and bass lines are assumed to have quarter notes only. The chord candidates were 12 major and 12 minor chords—that is, the elements in the set $\mathcal{C} = \{\mathrm{C}, \mathrm{C}^\sharp, \cdots, \mathrm{B}\} \times \{\mathrm{maj}, \mathrm{min}\}$. We also assumed the measure and key to be 4/4 and C major, respectively. In sum, the input was an m-bar chord progression $C = (c_0, \cdots, c_{m-1})$ $(c_i \in \mathcal{C})$, while the output was a bass line consisting of four quarter notes for each of m bars, for a total of $4m$ quarter notes, denoted by $B = (b_0, \cdots, b_{4m-1})$, in which each b_i is a MIDI note number.

2.2 Formulation with HMM

An HMM is a model in which an observed symbol is probabilistically emitted from a hidden state. In our case, we regarded given chords to be observed symbols and bass notes to be hidden states. Given a chord progression $C = (c_0, \cdots, c_{m-1})$, the observation $X = (x_0, \cdots, x_{4m-1})$ is defined so that $x_i = c_{[i/4]}$, in which [ ] is the floor function. Given $C = (\mathrm{C}, \mathrm{Am}, \mathrm{Dm}, \mathrm{G})$, for example, $X = (\mathrm{C}, \mathrm{C}, \mathrm{C}, \mathrm{C}, \mathrm{Am}, \mathrm{Am}, \mathrm{Am}, \mathrm{Am}, \mathrm{Dm}, \mathrm{Dm}, \mathrm{Dm}, \mathrm{Dm}, \mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{G})$.

We designed a set of hidden states $\mathcal{S}$ in three ways:

Method 1 The simplest, octave-ignored method
$\mathcal{S}$ was a set of pitch classes such that $\mathcal{S} = (0, 1, 2, \cdots, 11)$, where $0, 1, 2, \cdots, 11$ correspond to C, C$^\sharp$, D, $\cdots$, B, respectively.

Method 2 The simple, but non-octave-ignored method
In Method 2, $\mathcal{S}$ was a set of pitches in a specific pitch range such that $\mathcal{S} = (28, 29, 30, \cdots, 60)$, in which each integer represents a MIDI note number. This pitch range was determined so that it covered the pitches used in typical bass lines.

Method 3 The octave-ignored but metrical-position-considered method
$\mathcal{S} = (0, 1, 2, \cdots, 47)$ was a combination of pitch classes and metric positions. Specifically, each element s in $\mathcal{S}$ was calculated by $s = n + 12q$ where $n \in \{0, 1, 2, \cdots, 11\}$ is a pitch class and $q \in \{0, 1, 2, 3\}$ is a metric position. With that model, distribution of emission probability can be learned separately for each metric position. Because bass note selection obviously depends on its metric position—for example, the root note frequently occurs at the first beat in a bar but seldom at the last beat—we expected Method 3 to be superior to the others.

We did not consider the fourth possiblity, i.e., the non-octave-ignored metrical-position-considered method, because our dataset was too limited to learn such a large model.

Let $H = (h_0, \cdots, h_{4m-1})$ be a sequence of hidden states for a given observation $X = (x_0, \cdots, x_{4m-1})$. The following probabilities were learned from the dataset:

- the initial probabilities $\{P(h_0 = s) \mid s \in \mathcal{S}\}$,
- the emission probabilities $\{P(x_i = c | h_i = s) \mid c \in \mathcal{C}, s \in \mathcal{S}\}$, and
- the transition probabilities $\{P(h_{i+1} = s' | h_i = s) \mid s, s' \in \mathcal{S}\}$.

2.3 Algorithm for determining pitches

Given the chord progression $C = (c_0, \cdots, c_{m-1})$, we estimated the most likely sequence of hidden states $H = (h_0, \cdots, h_{4m-1})$ based on the HMM.

Because Methods 1 and 3 determined not pitches but pitch classes, we had to determine the octave for each bass note in order to determine the pitch. Let o_i be the octave—specifically, the MIDI note number of the octave's C—for the bass

note b_i. For Method 1, $b_i = o_i + h_i$, whereas for Method 3, $b_i = o_i + \mathrm{mod}(h_i, 12)$, in which mod is the modulo operation.

For the initial note, the lowest pitch within the specified pitch range (28 to 60) is selected, that is,

$$o_1 = \begin{cases} 24 \ (\mathrm{mod}(h_1, 12) \geq 4) \\ 36 \ (\mathrm{mod}(h_1, 12) \leq 3) \end{cases}$$

The octave of each other notes is determined so that the note is smoothly connected from the previous note, that is,

$$o_i = \begin{cases} \max(o_{i-1} + 12, 24) & (\mathrm{mod}(h_i, 12) - \mathrm{mod}(h_{i-1}, 12) < -5) \\ \min(o_{i-1} - 12, 48) & (\mathrm{mod}(h_i, 12) - \mathrm{mod}(h_{i-1}, 12) > 5) \\ o_{i-1} & (\text{otherwise}) \end{cases}$$

For Method 2, $b_i = h_i$ for every i.

3 Experiments

3.1 Dataset

We used data collected from "Jazz Bass Learning: 104 Examples Collections 1–3" [11–13] and the website Projazz Lab [14]. Of the bass lines collected and the transcriptions of chord progressions, ones with multiple chords in any bar were excluded as well as ones with bass notes beyond the specified bass pitch range (i.e., 28 to 60). We transposed all remaining bass lines to C major and divided them into four bars each. If a bass line contained non-quarter notes (e.g., a dotted quarter note plus an eighth note), we manually simplified it to a sequence of quarter notes. The total number of the four-bar bass lines collected was 206, 103 of which we used as training data, whereas we used the other 103 as test data.

3.2 Examples

The bass lines generated by Methods 1–3 for three chord progressions appear in Figs. 1–3, which show the following:

Fig. 1 In the bass lines generated with Methods 1 and 2, repetitions of the same notes appear in the second half. By contrast, the bass line generated by Method 3 is smooth overall, although a non-chord note appears in the first bar.

Fig. 2 The bass lines generated by Methods 1 and 2 are highly monotonous because the same notes appear repeatedly, whereas the one generated by Method 3 has many passing notes, which makes it smooth and melodious.

Fig. 3 Similarly to the other two examples, the bass lines generated by Methods 1 and 2 repeat the same notes, whereas the one generated by Method 3 has many passing notes that make it smoother.

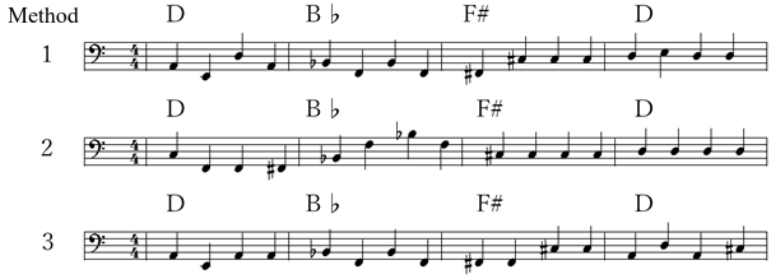


Fig. 1. Bass lines generated for D–B$^{\flat}$–F$^{\sharp}$–D

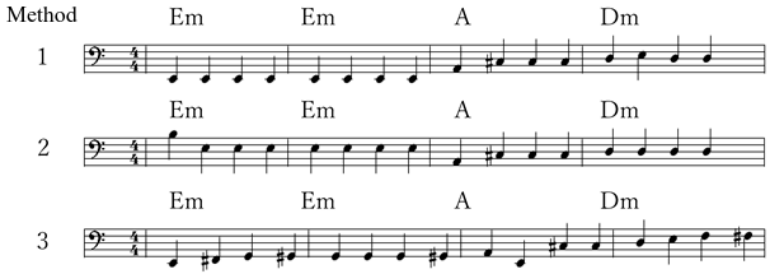


Fig. 2. Bass lines generated for Em–Em–A–Dm

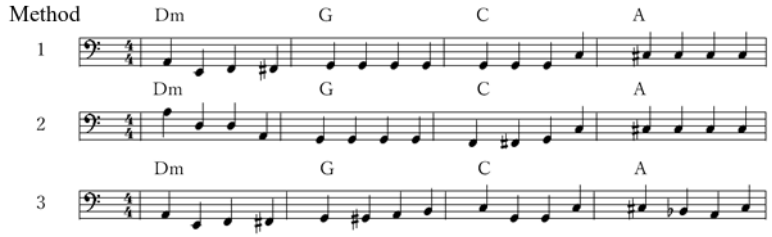


Fig. 3. Bass lines generated for Dm–F–C–A

3.3 Objective evaluation

We evaluated the bass lines generated with the 103 test data from two points of view. One evaluation involved note-wise comparison, in which we computed the rate of concordance with the ground truth (**C1**). If this rate is higher, the result can be considered better. The other evaluation is a statistics-based comparison, in which statistics such as **C2**–**C9** were computed from the bass lines generated and the ground truth. If those statistics are closer to the ones computed from the ground truth, the result can be considered better.

C1 Rate of concordance with the ground truth
C2 Rate of the root note of the chord at each bar
C3 Rate of the root note of the chord at the first beat of each bar
C4 Rate of the chord note at the first beat of each bar

Table 1. Results of the objective evaluation. (The bolded values are the best results, that is, the highest for **C1** and the closest to the ground truth for **C2** to **C9**).

	Method 1	Method 2	Method 3	Ground truth
C1	38.47%	35.32%	**41.75%**	—
C2	**54.79%**	56.55%	32.89%	51.33%
C3	53.88%	52.91%	**76.46%**	90.29%
C4	85.68%	81.07%	**96.36%**	98.05%
C5	11.10%	17.05%	**24.33%**	25.36%
C6	51.13%	**49.19%**	17.67%	41.88%
C7	31.59%	28.67%	**53.07%**	66.21%
C8	17.28%	22.14%	**29.26%**	43.04%
C9	4.20	5.40	**7.78**	8.68

C5 Rate of dissonant notes (notes with intervals of the minor 2nd from any chord notes)
C6 Rate of flat motions (here meaning a motion from a pitch to the same pitch)
C7 Rate of conjunct motions (motions with intervals of minor or major 2nd)
C8 Rate of distinct motions (motions with intervals of more than major 2nd)
C9 Number of pitch classes appearing in the bass line

The results listed in Table 1 can be summarized as follows:

C1 Method 3 showed the highest rates of concordance with the ground truth.
C2, C9 Method 3 generated bass lines containing various pitch classes (7.78) whereas Methods 1 and 2 generated bass lines in which the numbers of pitch classes were 4.2 and 5.4 on average, respectively. In addition, bass lines generated by Methods 1 and 2 contained many root notes (54.79% and 56.55%, respectively), as shown from **C2**. Of those results, bass lines generated by Methods 1 and 2 were more monotonous than those generated by Method 3.
C2, C3 Method 3 generated bass lines in which the first note at each bar was mostly the root note. In Methods 1 and 2, **C3**'s values were exceptionally close to **C2**'s values, because the distribution of the emission probabilities was common among all metric positions. By contrast, in Method 3, **C3**'s value was high (76.46%) even though **C2**'s value was low (32.89%), because that method prescribed separate distributions of emitted bass notes for each metric position.
C5 Method 3 generated bass lines with the highest rate of dissonant notes; however, that outcome was not problematic because the ground truth also had a similar rate of dissonant notes.
C6, C7, C8 Methods 1 and 2 generated bass lines with high rates of flat motions. By contrast, Method 3 generated bass lines with 53.07% of conjunct motions. This is a successful result since the high rate of distinct motions is a characteristic feature of walking basslines.

To summarize, Method 3 generated the most melodious bass lines, with a tradeoff of melodiousness and low dissonance. Method 1 mostly generated consonant bass

Table 2. Subjective evaluation results

	Method 1	Method 2	Method 3	Ground truth
S1	1.40	1.88	0.60	0.98
S2 (a)	3.52	3.00	3.70	3.96
S2 (b)	3.40	2.96	3.80	3.96
S2 (c)	3.54	3.16	3.82	4.10

lines, but half notes in the generated bass lines were root notes and had many flat motions. In contrast, the basslines generated by Method 3 included many non-root notes and moderate conjunct motions.

3.4 Subjective evaluation

We asked an expert bassist with 25 years of experience in playing jazz bass to evaluate the bass lines generated by the three methods. We selected 50 chord progressions at random from the ones used in the objective evaluation and prepared four bass lines (i.e., Methods 1–3 and the ground truth) for each chord progression. We gave the bassist the scores and MP3 data of the prepared bass lines and asked him to evaluate them in the following ways:

S1 Marking musically inapporopriate notes in the scores—the fewer, the better.
S2 Rating the bass lines from the following criteria on a scale of 1 to 5, in which higher ratings indicated better bass lines:
(a) Overall quality
(b) Overall smoothness
(c) Congruency with the chord progression

The results of the expert's evaluation appear in Table 2 and can be summarized as follows:

S1 The number of musically inappropriate notes was minimal for Method 3 among the three methods and ground truth, for two possible reasons. One is that some bass lines of the ground truth consisted primarily of root notes to allow novice players to play. The other is that some bass lines may have been made too simple by the simplification mentioned in Section 3.1.
S2 For all of (a) to (c), Method 3 obtained scores between 3.7 and 3.9 and was superior to the other methods, even though it was slightly worse than the ground truth.

4 Conclusion

In this paper, we have proposed a method of generating walking bass lines for jazz using an HMM. By designing hidden states consisting of combinations of pitch classes and metric positions, the model successfully learned different tendencies of bass notes at different metric positions. As a result, the bass lines generated contained passing notes and were thus musically flowing.

Our study involved several limitations. For one, we considered four-bar bass lines only. For longer musical pieces, a bass line should vary along the progress of the music. In the first verse and another verse later, for example, the bassist may play different bass lines even if the melody and chord progression remains the same. In response, we need to consider the relationship of bass lines and musical context in order to generate appropriate bass lines for long pieces. Moreover, advanced bassists often add ornamentation to bass lines, which we also did not consider in our study. Also, it could be more useful to adapt generated bass lines to the user's performing skill.

References

1. K. Ebcioglu. An expert system for harmonizing four-part chorales. *Computer Music Journal*, 12(3):43–51, 1988.
2. Hermann Hild, Johannes Feulner, and Wolfram Menzel. Harmonet: A neural net for harmonizing chorales in the style of J.S. Bach. In *Advances in neural information processing systems*, pages 267–274, 1992.
3. T. Kawakami, M. Nakai, H. Shimodaira, and S. Sagayama. Hidden Markov model applied to automatic harmonization of given melodies. In *IPSJ SIG Notes*, 99-MUS-34, pages 59–66, 2000. in Japanese.
4. Somnuk Phon-Amnuaisuk, Alan Smaill, and Geraint Wiggins. Chorale harmonization: A view from a search control perspective. *Journal of New Music Research*, 35(4):279–305, 2006.
5. Jan Buys and Brink van der Merwe. Chorale harmonization with weighted finite-state transducers. In *Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, pages 95–101. PRASA South Africa, 2012.
6. Syunpei Suzuki and Tetsuro Kitahara. Four-part harmonization using Bayesian networks: Pros and cons of introducing chord nodes. *Journal of New Music Research*, 43(3):331–353, 2014.
7. Rui Dias and Carlos Guedes. A contour-based jazz walking bass generator. In *Sound and Music Computing Conference (SMC 2013)*, pages 305–308, 2013.
8. Kanae Kunimatsu, Yu Ishikawa, Masami Takata, and Kazuki Joe. A music composition model with genetic programming —a case study of chord progression and bassline. In *Proceedings of Int'l Conf. Par. and Dist. Proc. Tech. and Appl. (PDPTA 2015)*, pages 256–262, 2015.
9. Geber L. Ramalho, Pierre-Yves Rolland, and Jean-Gabriel Ganascia. An artificially intelligent jazz performer. *Journal of New Music Research*, 28(2):105–129, 1999.
10. Pere Calopa Piedra. Bassline generation agent based on knowledge and context. B.Sc. Thesis, Universitat Pompeu Fabra, 2015.
11. Shintaro Nakamura. *Jazz Bass Learning: 104 Examples Collection 1.* Saber Incorporated, 2017. in Japanese.
12. Shintaro Nakamura. *Jazz Bass Learning: 104 Examples Collection 2.* Saber Incorporated, 2017. in Japanese.
13. Shintaro Nakamura. *Jazz Bass Learning: 104 Examples Collection 3.* Saber Incorporated, 2017. in Japanese.
14. http://www.projazzlab.com/study-tools.

Programming in style with *bach*

Andrea Agostini[1], Daniele Ghisi[2], and Jean-Louis Giavitto[3]

[1] Conservatory of Turin
[2] Conservatory of Genoa
[3] CNRS, STMS – IRCAM, Sorbonne University

Abstract. Different programming systems for computer music are based upon seemingly similar, but profoundly different, programming paradigms. In this paper, we shall discuss some of them, with particular reference to computer-aided composition systems and Max. We shall subsequently show how the *bach* library can support different programming styles within Max, improving the expression, the readability and the maintainance of complex algorithms. In particular, the forthcoming version of *bach* introduces *bell*, a small textual programming language embedded in Max and specifically designed to facilite programming tasks related to manipulation of symbolic musical material.

Keywords: Programming paradigms, computer-aided composition, Max, *bach*, *bell*

1 Introduction

In spite of the way it is advertised, its own Turing-completeness and the sheer amount and complexity of things that have been done with it, programming in Max is difficult. Whereas setting up simple interactive processes with rich graphical interfaces may be immediate, it has been long observed that implementing nontrivial algorithms is far from straightforward, and the resulting programs are often very difficult to analyse, maintain and debug.

Several other popular programming languages and environments for computer music, such as OpenMusic [1], PWGL [10] and Faust [11], share with Max a superficially similar, but profoundly different, dataflow programming paradigm, which makes them better suited for 'real' programming and less for setting up highly interactive and responsive systems. This is reflected in the types of artistic practices these systems are typically used for, and mirrors the oft-discussed rift between composition- and performance-oriented tools in computer music.

We are convinced that this rift is by no means necessary or natural, and, on the contrary, has proven problematic with respect to a wide array of practices lying somehow between the two categories, such as extemporaneous, 'intuitionistic' approaches to composition (including, but not limited to, improvisation), sound-based and multimedia installations, live coding and more.

In this paper, we shall investigate this divide and its reasons from the point of view of computational models, and consider how it can be bridged, or at least narrowed, through the use of the *bach* package for Max [3].

2 Dataflow computational models

The concept of data flow is an old one, dating back at least to [5], which first introduced the idea of independent computational modules communicating by sending data (discrete items) among directed links. Over the years, many kinds of dataflow computation models have been developed. In this section, we shall review some of them and how they apply to different languages and software systems for computer music.

2.1 The pipelined and functional dataflow and the Kahn principle.

Several computer music languages and systems (such as Reaktor, OpenMusic and Faust, but also Reaper, Live, ProTools and various software synthesisers) are based upon the pipelined dataflow model. This means that programs written in those systems have the following features, strictly linked to one another:

- Programs are represented as directed graphs; each node of the graph implements an abstract process consuming data on its input links and producing data on its output links; the links are the only interactions between the processes; the sequence of data traversing a link is called a *stream*.
- Processes have *referential transparency*: we can always replace every variable and function by its definition, and a function called twice on the same data will always return the same result.
- The resulting programming style is *declarative*: programmers only specify the properties of the objects they want to build, rather than the way to build them.
- Programs are mathematical objects, and can be treated as such. It has been proven by Gilles Kahn [9] that a pipelined dataflow program is equivalent to a set of equations, taking the form of a fixed point equation. What is known now as the Kahn Principle states that the stream associated to each edge of the dataflow graph is the solution of the previous set of equations. As a consequence, algebraic reasoning on the operational properties of programs is possible and useful.
- Just like in the process of solving an equation, there is no notion of temporal sequencing of actions, but rather of algebraic relations between parts of the equation. Therefore, the program graph is acyclic (as feedback loops are only meaningful if they establish a temporal delay), and an input link can only accept one single incoming graph edge.

All this being considered, whether to write a functional dataflow program as a graphical patch or as a set of equations specified textually is a matter of taste. Faust is an example of a textual, functional dataflow programming language in which a program is a set of equations.

Although programs are not explicitly expressed as equations, OpenMusic and PWGL are essentially based upon the same assumptions as Faust.[4] How-

[4] It should be remarked that the computational model of OpenMusic and PWGL is not as pure as described here, since it includes imperative traits like storing and retrieving mutable states through variables.

ever, evaluation in Faust is driven by the availability of the data, that is, it happens whenever data enter the nodes. On the contrary, the evaluation process of OpenMusic and PWGL is demand-driven, that is, the user requests a result to the bottom node of the graph, which in turns requests values to the nodes connected to its input links, and so on.

2.2 Asynchronous, non-functional pipelined dataflow: Max

Max implements two different dataflow systems, respectively devoted to audio signals and control messages.[5] The former is a relatively simple case of synchronous pipelined dataflow, whose functional nature is somewhat less explicit than that of Faust but not too different from it: in fact, the functional dataflow view fits very well with the audio graph representation of signal processing. Our discussion will only focus on the latter and its significantly different paradigm. In what follows, we shall assume in the reader a basic, practical knowledge of Max, and only review some fundamental concepts when needed.

A Max patch can be seen as a set of nodes working asynchronously with respect to each other: if, when and how each module 'fires' depends on the data processed, and, generally speaking, only one message can traverse the patch at any given time. This means that nodes with more than one input link must have mechanisms for storing data for later use. This is accomplished through the so-called 'hot' and 'cold' inlets (that is, input links in the Max jargon): when a hot inlet receives a message, it performs its computation and delivers the result; but when a message is received in a cold one, it gets stored for later use and nothing else happens. Most Max objects have at least one hot inlet, and many have one or more cold inlets.

This structure, which actually involves many other details and is not without exceptions, has some profound consequences:

- Objects have mutable states, which may change over time in response to a single evaluation request (see, e.g., the 'cold' inlet of any arithmetical operator).
- Objects have no referential transparency. The order of messages on a link is not enough to determine the global behavior of a patch: the precise timestamp of these messages is semantically meaningful. Altering the order in which data are sent from a node to others in response to a single piece of incoming data (that is, to a single computation request) may change the performed computation.
- Multiple links ('cords' in the Max jargon) can be connected to a single inlet: as data are always transmitted sequentially, this means that the inlet will receive data from its incoming cords one after another, and act consequently.

[5] Most of the general principles described here also apply to Max's sibling system, Pd, which we shall not discuss as the *bach* library is currently not available for it.

2.3 Pros and cons of different computational models

Max's computational model is motivated by the fact that, unlike the other systems described above, it was not conceived as a programming language but, in its own creator's words [12], as a *musical instrument*. With respect to this end, Max has the merit of being extremely economical in terms of its basic principles and quite adaptable to very different use cases.

On the other hand, as hinted at above, representing nontrivial algorithms in Max is often more complicated than with other systems. Two of the authors became painfully aware of this complicatedness while working at the *cage* package [2], which implements a comprehensive set of typical computer-aided composition operations. *cage* is entirely composed of abstractions, and during its development the shortcomings of Max programming became so evident that the seeds for the work presented in this article were planted.

The reasons for this difficulty are multiple, and include the following:

- The greater freedom Max grants in building the program graph easily leads to far more intricate patches than functional dataflow models, with spaghetti connections that can grow very hard to analyse.
- Typical Max patches often have their state distributed through many objects whose main, individual purpose is not data storage.
- Max lacks, or implements in quite idiosyncratic ways, some concepts that are ubiquitous in modern programming languages, such as complex, hierarchical data structures, data encapsulation, functions and parametrization of a process through other processes.

On the other hand, Max allows to incorporate, on top of its basic paradigm, traits reminiscent of various programming styles, such as imperative, object-oriented and functional. Moreover, it includes various objects enclosing entire language interpreters, thus allowing textual code in various languages to be embedded in a patch.

These features may prove useful when nontrivial processes have to be implemented, as is the case when working in contexts like algorithmic and computer-aided composition. Whereas Max was not conceived with these specific applications in mind, it quickly became clear that it could be a valuable environment for them, and several projects have been developed in this sense [14,13,7]. We shall focus on one of them, the *bach* package, which has been conceived and maintained by two of the authors.

2.4 The *bach* package

The *bach* package[6] for Max is a freely available library of more than 200 modules aimed at augmenting Max with advanced capabilities of symbolic musical representation. At its forefront are two objects called `bach.roll` and `bach.score`, capable of displaying, editing and playing back musical scores composed of

[6] `www.bachproject.net`

both traditional notation and arbitrary time-based data, such as parameters for sound synthesis and processing, textual or graphical performance instructions, file paths and more.[7]

One of the main focuses of *bach* is algorithmic generation and manipulation of such scores. To this end, *bach* implements in Max a tree data structure called *llll* (an acronym for Lisp-like linked list), meant to represent arbitrary data including whole augmented scores. *bach* objects and abstractions exchange *llll*s with each other, rather than regular Max messages, and their majority is devoted to performing typical list operations such as reversal, rotation, search, transposition, sorting and so on.

Generally speaking, *bach* objects abide by the overall design principles and conventions of Max, but it should be remarked that, whereas standard Max objects can control the flow of *llll*s in a patcher just like they do with regular Max messages, they cannot access their contents unless *llll*s are explicitly converted into a Max-readable format, which on the other hand has other limitations (for a detailed explanation, see [3]). Thus, *bach* contains a large number of objects that somehow extend to *llll*s the functionalities of standard Max objects. For example, whereas the `zl.rev` object reverses a plain Max list, the `bach.rev` object reverses an *llll* by taking into account all the branches of the tree, each of which can be reversed as well or not according to specific settings. Whereas it is possible to convert an *llll* into Max format and reverse it with `zl.rev`, in general the result will not be semantically and syntactically correct.

Since its beginnings, *bach* has been strongly influenced by and related to a number of other existing projects: for an overview of at least some of them, see [3]. The synthesis of different approaches that lies at the very basis of the conception itself of *bach* has been validated by a large community of users, who have developed many artistic and research projects in several domains[8], as well as the fact that it provides the foundation for the *cage* and *dada*[9] libraries [8].

In the following sections, we shall review a few programming styles and approaches and see how *bach* can be helpful with adopting them in Max: namely, we shall show how some fundamentally imperative, functional and objected-oriented traits of Max can be leveraged through the use of specific *bach* objects and design patterns; moreover, we shall discuss a recent addition to *bach*, that is, a multi-paradigm programming language called *bell* and meant to facilitate the expression of complex algorithms for manipulating *llll*s.

[7] `bach.roll` and `bach.score` differ in that the former represents time proportionally, whereas the latter implements a traditional representation of time, with tempi, metri, measures and relative temporal units such as quarter notes, tuplets and so on.

[8] The website of *bach* showcases some interesting works that have been developed with the library, mostly by people independent of its developers.

[9] The *dada* library contains interactive two-dimensional interfaces for real-time symbolic generation and dataset exploration, embracing a graphic, ludic, explorative approach to music composition.

3 Different programming styles and approaches in Max

3.1 Imperative approach

It has been observed that Max is essentially an imperative system in disguise [6]: as stated before, any nontrivial program in Max requires to take care of states and the order of operations, and analysing even a moderately complex patch can only be done by following the flow of data and the evolution of states over time.

It is possible to make this imperative style more explicit by adopting some good practices, such as widely using specific objects, such as `trigger` and `bangbang`, that can help with keeping the evaluation order under control. Moreover, Max contains two objects whose only purpose is holding data associated with a name: `value` and `pv` (for 'private value'), whose role can be seen as corresponding to that of variables in traditional imperative programming languages. Each instance of those objects has a name, and every time it receives a piece of information it retains and shares it with all the other objects with the same name. It is subsequently possible retrieve the stored data from any of them. The `value` and `pv` modules differ in their scope: the former's is global, that is, data are shared through all the open patches in the Max session, whereas the latter's is local, in that data are only shared within the same patcher or its subpatchers. By combining `value` and `pv` with the aforementioned sequencing objects, it is possible to use Max in a much more readable, essentially imperative programming style.

bach implements its own variants of these objects, respectively named `bach.value` and `bach.pv`. Besides dealing correctly with *llll*s, they can open a text editing window if double-clicked, allowing to view and modify the data they hold. Moreover, *bach* contains an object called `bach.shelf`, which acts as a container of an arbitrarily large set of *llll*s, each associated to a unique name. `bach.shelf` objects can be themselves named, thus defining namespaces: this means that *llll*s associated to a name within one named `bach.shelf` object will be shared only with other `bach.shelf` objects with the same name. Although still somewhat crude (it might be interesting, for example, allowing non-global namespaces), this is a way to improve data localization and data encapsulation, and reduce the proliferation of storage objects in complex scenarios.

3.2 Object-oriented approach

The fact that a Max program is built of independent blocks responding to messages they send to each other in consequence of callbacks triggered by events gives it a strong object-oriented flavour, and the Smalltalk influence is both apparent and declared. At a lower level, in fact, each Max object in a patch is an instance of a specific class, with member variables containing the object's state and methods roughly corresponding to the messages it accepts for modifying and/or querying the state.

The two main *bach* editors, `bach.roll` and `bach.score`, comply with this object-oriented approach. However, a distinction can be made about the kinds

of messages they accepts: some control and query the object's appearance (background color, zoom level, etc.), whereas others are dedicated to the direct management of the editor's content. This distinction between the two kinds of messages is explicit in the syntax of the messages they receive.

Messages dealing with the editor's contents enable the creation, the edition and the deletion of individual notation items, such as a single measure or a single note. These messages can actually be seen as methods of the items themselves, which are arranged according to a precise hierarchy and share a certain number of common properties (such as having a symbolic name, being selectable, etc.).

In fact, there are several ways to modify a score. One of the simplest involves dumping its parameters from some outlets, modifying them via appropriate Max and *bach* modules, and feeding the result into a different editor object.

In contrast, one can send direct messages to the editor, asking for specific elements of the score to be created or modified through the so-called *bach in-place syntax*, with no output from the object outlets (unless explicitly requested). Modifications are immediately performed and the score is updated (see Fig. 1). This mechanism is strongly inspired by an object-oriented approach: first, references to the notation items to be modified are acquired via a selection mechanism, and then messages are sent to them. For example, a set of notes can be selected graphically, or through a query in the form of a message such as `sel note if voice == 2 and pitch % C1 == F#0`. After this, those notes can be modified by means of messages such as `duration = velocity * 10`.

In fact, this kind of approach allows much more complex operations than the ones described here, as there are many classes of notation items, each having a large number of properties and related messages. In spite of the richness of the data it can manipulate, though, the in-place syntax is not very flexible, but there are plans to extend it through the *bell* language (see below).

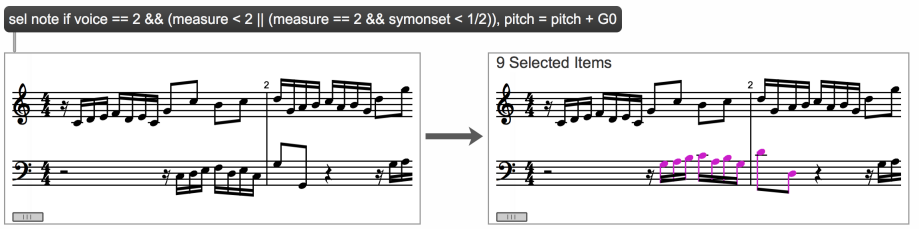


Fig. 1. A very simple example of in-place modification: notes belonging to the second voice and whose onset lies before the middle of the second measure are selected and transposed up a perfect fifth (the image shows both the state of the score before and after the click on the message).

3.3 Functional approach

Max shares some similarities with functional languages, mostly by handling values through a variety of nodes implementing functions on these values. It is then possible to build patches that somehow behave functionally, and whose appearance is extremely similar to that of equivalent ones in a functional graphical system such as PWGL. *bach* extends the functional traits of Max in a few areas.

As hinted at before, it implements the *llll*, a tree data type quite similar to a Lisp list, and provides a large number of modules for dealing with *llls*. Although, of course, list operators are not inherently functional, they are quite customary in functional languages, and the corresponding *bach* objects can be connected in a way corresponding to the composition of list functions in functional languages such as Lisp or Haskell.

Secondly, generalized versions of functions such as *sort* and *find* require some way to specify, respectively, a custom ordering or an arbitrary search criterion. In several languages, these generalized functions are conveniently implemented as higher-order functions, i.e., functions taking other functions as arguments. This requires to handle functions like ordinary data. A Max patcher lacks the concept of function, but several *bach* objects implement a design pattern called the *lambda loop* (see Fig. 2), whose role is somehow akin to that of higher-order functions.

A lambda loop is a patching configuration in which one or more dedicated outlets of a module output data iteratively to a patch section, which must calculate a result (either a modification of the original data, or some sort of return value) and return it to a dedicated inlet of the starting object [3].

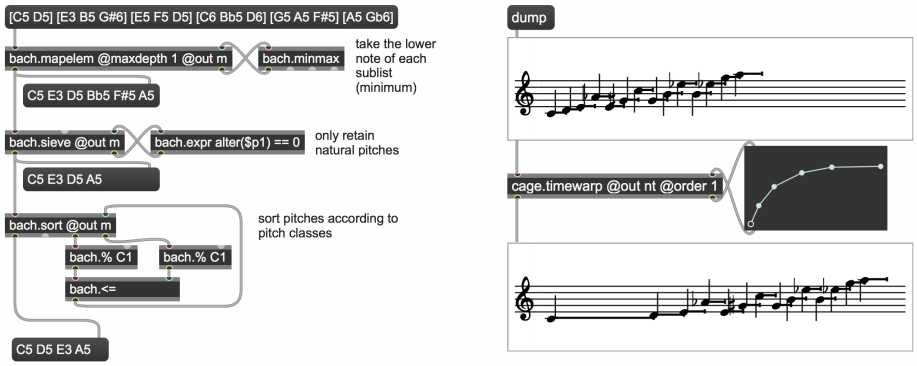


Fig. 2. The cross-connected and loop-connected patch cords attached to *bach.mapelem*, *bach.sieve*, *bach.sort* and *cage.timewarp* modules form several instances of the so-called lambda loop. The left-side example should be straightforward. In the right-side example, the temporal distribution of events in a musical score is altered through the provided transfer function, with time on the X axis and speed on the Y axis. At a superficial level, patches like these appear to be quite similar to how the same processes might be implemented in a functional dataflow system.

Lambda loops are used by some *bach* modules directly inspired by functional programming practices, such as `bach.mapelem` (performing a map operation) and `bach.reduce` (recursively applying a binary function on elements); all these modules can be helpful to translate programs conceived functionally into Max patches. The number of modules taking advantage of this design pattern is, however, much larger, and include basic operators such as `bach.sieve` (only letting some elements through) and `bach.sort` (performing sort operations), but also advanced tools such as `bach.constraints` (solving constraint satisfaction problems) as well as some of the modules in the *cage* package.

4 Textual coding

The approaches described so far are based on the idea that individual objects carry out elementary operations, and they are connected graphically so as to build complex behaviors.

A different, but not incompatible, point of view is embedding an algorithm, even a potentially complex one, into a single object by means of textual coding, and subsequently insert it into a patch. In graphical, Lisp-based systems such as OpenMusic and PWGL, this is easily accomplished by inserting graph boxes containing Lisp code in the patcher. Whereas it is possible to embed in Max textual code written into various programming languages including C, Lua, Java and JavaScript, we feel that none of those language bindings provides the ease and directness of embedding of a Lisp code box in OM or PWGL.

On the other hand, Max contains a family of objects, namely `expr`, `vexpr` and `if`, that allow defining textually mathematical expressions and simple conditionals which might otherwise require fairly complicated constellations of objects in a patch. *bach* adds another member to the family, called `bach.expr`, allowing to define mathematical expressions to be performed point-wise on *llll*s.

Whereas the `expr` family syntax is not a full-fledged programming language, it can be seen as the basis for one. We therefore decided to include in the latest release of *bach* a new object to the family, called `bach.eval`, implementing a new, simple programming language conceived with a few, conceptually simple points in mind:

- Turing-complete, functional syntax, in which all the language constructs return values, but also including imperative traits such as sequences, variables and loops.
- Full downward compatibility with the `expr` family.
- Inclusion of list operators on *llll*s respecting, as far as possible, the conventions and naming of the corresponding *bach* objects.
- Implicit concatenation of elements into *llll*s, meaning that by simply juxtaposing values (be they literals, or the result of calculations) they are packed together into an *llll*. In this way, a program can be seen as an *llll* intermingled with calculations, not unlike what happens by combining the `quote` operator and `unquote` macro in Lisp.

– Maximum ease of embedding of the object into a Max patcher, with, among the other things, no need for explicit management of inlets and outlets.

The resulting language is called *bell* (standing for *bach evaluation language for llls*, but also paying homage to the historic Bell Labs). A detailed description of its syntax can be found in [4], whereas, for the scope of this article, a few examples should suffice (see Fig. 3, 4 and 5).

bell code can be typed in the `bach.eval` object box or into a dedicated text editor window, loaded from a text file and even passed dynamically to the host object via Max messages.

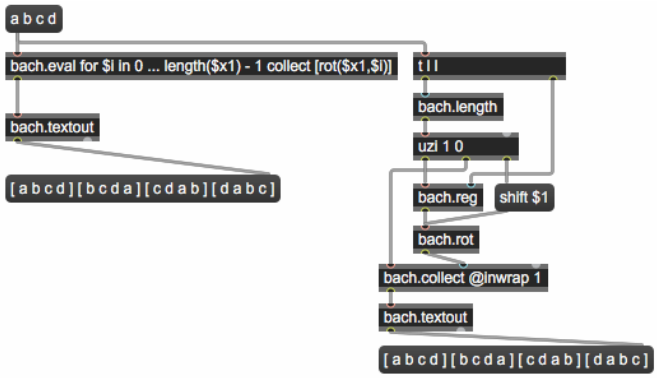


Fig. 3. A comparison between an *llll* manipulation process described through a snippet of *bell* code (in the `bach.eval` object box) and the corresponding implementation within the standard graphical dataflow paradigm of Max. The code should be mostly straightforward for readers familiar with the *bach* library and a textual programming language such as Python, considering that the `[ ... ]` paired operator encloses one or more elements into a sublist, according to the general syntax of *llls*.

The intended usage paradigm of `bach.eval` is similar to that of the expr family: `bach.eval` objects are meant to carry out relatively simple computational tasks, and to be sprinkled around the patcher among regular *bach* and Max objects taking care of the UI, MIDI, DSP, event scheduling and so on.

Snippets of *bell* language can also be passed to other objects as well for fine-tuning their behavior, as a replacement for lambda loops. Moreover, an intended (albeit not straightforward) development is to allow `bach.score` and `bach.roll` to be scripted in *bell*, thus allowing far more complex interactions than what is already possible through the syntax described above.

5 Conclusions and future work

We have presented some historical and theoretical background about the computational models of Max and other related programming languages and envi-

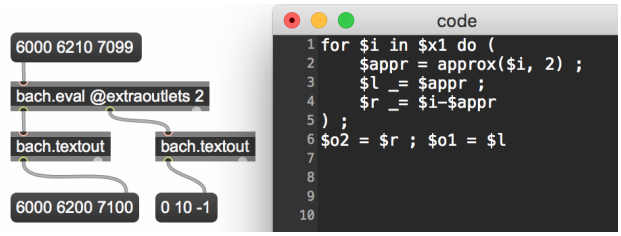


Fig. 4. A snippet of *bell* code approximating a list of midicents to the nearest semitone, and returning the distances from the semitone grid from a different outlet. Here, the code has been typed in a separate text editing window (shown on the right). The `$o1` and `$o2` pseudovariables assign results to the extra outlets declared in the `bach.eval` object box. The main, rightmost outlet returning the actual result of the computation (which, in this example, is the last term of the sequence defined by the `;` operators, that is, the value of the `$l` variable as passed to the first extra outlet) is left unused here. The language has several other features not shown here, including named and anonymous user-defined functions with a rich calling mechanism.

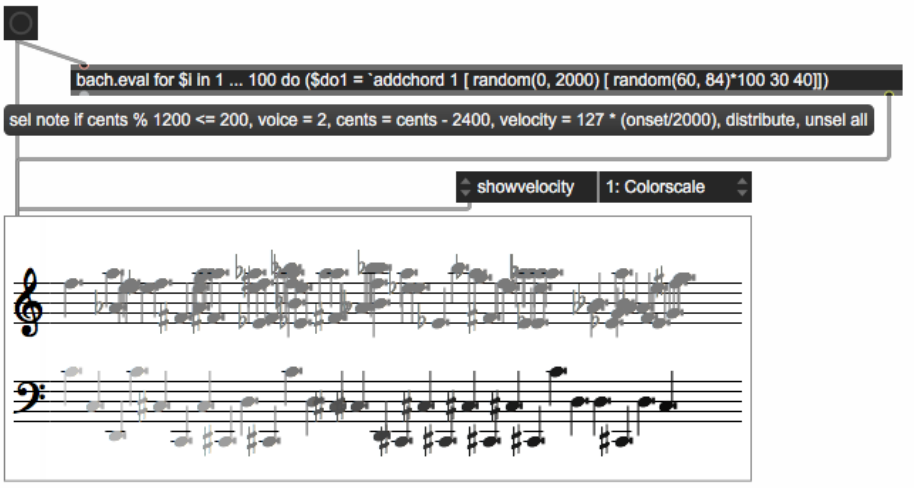


Fig. 5. An example of usage of *bell* in combination with `bach.roll`'s in-place syntax: 100 notes are generated in the first voice with random onsets (between 0 and 2 seconds) and random pitches (between middle C and the C two octaves above, on a tempered semitonal grid); then all C's, C♯'s and D's are selected (i.e. notes whose remainder modulo 1200 is less than or equal to 200), assigned to the second voice, transposed two octaves below, remodulated with a velocity crescendo and distributed equally in time.

ronments, and subsequently described how the *bach* library can be helpful with writing clear and maintainable programs, through some specific features aimed at implementing different programming approaches and styles on top of it.

Overall, we think that time is ripe for advocating the adoption of more structured and theoretically grounded approaches to working with this successful and widely used tool. We hope that this article may be a step in that direction: further steps should involve, on the one hand, an actual survey of real-life use cases, possibly with the involvement of the community of *bach* users; and, on the other hand, a more precise and organic formalisation of good and scalable programming practices in Max, which might prove quite different from the ones typical of more traditional programming languages.

References

1. C. Agon. *OpenMusic : Un langage visuel pour la composition musicale assistée par ordinateur.* PhD thesis, University of Paris 6, 1998.
2. A. Agostini, E. Daubresse, and D. Ghisi. *cage*: a High-Level Library for Real-Time Computer-Aided Composition. In *Proceedings of the International Computer Music Conference*, Athens, Greece, 2014.
3. A. Agostini and D. Ghisi. A Max Library for Musical Notation and Computer-Aided Composition. *Computer Music Journal*, 39(2):11–27, 2015/10/03 2015.
4. A. Agostini and J. Giavitto. *bell*, a textual language for the bach library. In *Proceedings of the International Computer Music Conference (to appear)*, New York, USA, 2019.
5. M. E. Conway. Design of a separable transition-diagram compiler. *Communication of the ACM*, 6(7):396–408, 1963.
6. P. Desain et al. Putting Max in Perspective. *Computer Music Journal*, 17(2):3–11, 1992.
7. N. Didkovsky and G. Hajdu. Maxscore: Music Notation in Max/MSP. In *Proceedings of the International Computer Music Conference*, 2008.
8. D. Ghisi and A. Agostini. Extending bach: A family of libraries for real-time computer-assisted composition in max. *Journal of New Music Research*, 46(1):34–53, 2017.
9. G. Kahn. The semantics of a simple language for parallel programming. In *proceedings of IFIP Congress'74*, pages 471–475, 1974.
10. M. Laurson and M. Kuuskankare. PWGL: A Novel Visual Language based on Common Lisp, CLOS and OpenGL. In *Proceedings of International Computer Music Conference*, pages 142–145, Gothenburg, Sweden, 2002.
11. Y. Orlarey, D. Fober, and S. Letz. Faust: an efficient functional approach to dsp programming. *New Computational Paradigms for Computer Music*, 290:14, 2009.
12. M. Puckette. Max at seventeen. *Computer Music Journal*, 26(4):31–43, 2002.
13. S. Scholl. *Musik — Raum — Technik. Zur Entwicklung und Anwendung der graphischen Programmierumgebung "Max"*, chapter Karlheinz Essls RTC-lib, pages 102–107. Transcript Verlag, 2014.
14. T. Winkler. *Composing Interactive Music: Techniques and Ideas Using Max.* The MIT Press, 1998.

Method and System for Aligning Audio Description to a Live Musical Theater Performance

Dirk Vander Wilt and Morwaread Mary Farbood

New York University
[dirk.vanderwilt,mfarbood]@nyu.edu

Abstract. Audio description, an accessibility service used by blind or visually impaired individuals, provides spoken descriptions of visual content. This accommodation allows those with low or no vision the ability to access information that sighted people obtain visually. In this paper a method for deploying pre-recorded audio description in a live musical theater environment is presented. This method uses a reference recording and an online time warping algorithm to align audio descriptions with live performances. A software implementation that is integrated into an existing theatrical workflow is also described. This system is used in two evaluation experiments that show the method successfully aligns multiple recordings of works of musical theater in order to automatically trigger pre-recorded, descriptive audio in real time.

Keywords: audio description, blind, visually impaired, accessibility, disability, musical theater, time warping

1 Introduction

Audio description (AD) is an accommodation used by people who are visually impaired or blind. It is a spoken description of the visual elements of an accompanying work, providing an accessible alternative to obtaining information that sighted individuals may obtain visually. Users of AD should be able to ascertain with audio what a sighted person at the same time and place may ascertain with vision. At live theatrical events that provide AD, such as some Broadway musicals, patrons are either provided with a wireless audio receiving device or are asked to download a software application onto their smartphone, so the transmission of the AD will not disrupt other theatergoers. For an overview on the art and practice of AD, see Fryer [6] and Snyder [14].

Live theatrical events pose an interesting problem for AD services. Like fixed media, the descriptions must be timed appropriately so as not to disrupt other aural aspects of the show (musical numbers, dialogue, etc.) [6, 14]. However, since repeated live performances are by design never identical, a fixed AD track cannot be aligned in advance. In live situations, either a professional audio describer describes the visual elements in real time, or a system is created to allow pre-recorded AD to be triggered [9]. Developing and deploying this type of service

is expensive and time-consuming. A recent study showed that media producers view AD as "a costly service with no revenue potential." Creating audio description for a 30-minute television show with 24 cues may cost between $698 and $1,462, depending on how the description is produced [12]. According to Szarkowska [15], "A lengthy preparation process and high production costs are among the greatest obstacles to the wider availability of audio description."

This paper proposes an inexpensive and novel method to trigger AD for a live theatrical performance by only using audio obtained from a show's previous performance. The process described in this paper warps the audio from the live show in real time to a reference recording using an established online time warping algorithm [5]. This method is a step towards being able to reduce the cost of deploying live theatrical audio description, thus making it more available to visually impaired people.

2 Method

Live audio-to-audio alignment has been used successfully in music score following [1, 2] and audio-transcription alignment [8] tasks. In this implementation, a reference recording is aligned to live input using an online time warping algorithm [5]. During the live alignment, descriptive audio tracks based on their pre-aligned position in the reference recording are also aligned and played back for blind and visually-impaired audience members.

First, a performance of the entire production is recorded in advance. Relevant features from that audio are then extracted and stored in frames of vectors. A second audio track is also created, containing only the AD which aligns to that recorded performance. The descriptive track is broken up into multiple smaller tracks such that one sub-track contains the audio for a single described event within the recorded performance, and the points where each sub-track should be triggered are marked. Once the marks, descriptive track, and extracted features of the reference recording are obtained, the live alignment may begin.

2.1 Audio Features

In the system described here, Mel-frequency cepstrum coefficients (MFCCs) are extracted from both the live input and reference recording. MFCCs are a well-established set of features used when analyzing audio signals for human speech recognition and music information retrieval applications. Both uses are applicable here since theatrical productions often contain both speech (dialogue) and music (showtunes). The coefficients are derived from the Mel scale, which captures patterns audible to the human ear. Although the system does not recognize speech explicitly, it uses MFCCs to compare different (but similar) samples of speech and music patterns. The code implemented to extract MFCCs here was based on [7].

Starting with audio at a sampling of 8 kHz, the MFCC extraction process begins with the application of a pre-emphasis filter so that the higher frequencies

of the signal have greater energy. The signal is then segmented into frames, and a Hamming window and FFT are applied to each frame. The results are filtered through a Mel filterbank with 40 filters, which is where the raw frequency data gets "refined" to frequencies based on the Mel scale. A discrete cosine transform (DCT) is performed on the log-amplitudes of the filter output, resulting in 13 coefficients which are the MFCCs for that frame. To account for change over time, a first-order difference of each coefficient from the previous frame is appended to the 13 current coefficients, making the total number of coefficients used for analysis 26 per frame. Given the real-time nature of this system, the features must be extractable in less time than it takes the corresponding audio to play out in real time. In this case, the system extracts one MFCC feature vector at a frame length of 100 ms and a hop size of 40 ms. For a description of MFCC feature extraction, see [10, 8, 11].

2.2 Online Time Warping

Dynamic Time Warping (DTW) uses dynamic programming to recursively compare two time series U and V of lengths m and n. The output is an m- by- n matrix D where each element $D(x, y)$ is a cumulative (in this case Euclidean) distance between $U(x)$ and $V(y)$. The value of each cell is the cumulative "path cost" from $D(1, 1)$ up to that point [13]:

$$D(i,j) = d(i,j) + min \left\{ \begin{array}{l} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{array} \right\} \quad (1)$$

Every cell in the matrix is calculated by obtaining the distance between $U(i)$ and $V(j)$, and adding it to the least of one of three adjacent (previous) cells. In this way, the cumulative path cost between $D(1, 1)$ and the current cell is determined. The smaller the cumulative cost, the better the match up to that point. When the whole matrix is calculated through $D(m, n)$, backtracking the smallest distance back to $D(1, 1)$ will be the warp path which relates the closest points of U to V. Unfortunately, this algorithm requires both series to be fully available in advance and has a running time of $O(N^2)$, making it unsuitable for live performance tracking.

This online time warping algorithm, developed by Dixon [5], requires only one of the series to be available in advance, while the other series may be obtained in real time (V is known fully in advance, and U is only partially known, but increases as new live input is received). The algorithm outputs a similar matrix D, but it builds as the input is received, one row and/or column at a time, and only in the forward direction. Plus, it is only able to estimate the total size of the resulting matrix, so it is instead bounded by a constant, which is determined in advance. Thus, it does not have the advantage of being able to backtrack from future input.

In online time warping, whenever a new frame of input is received in real time as $U(t)$, where t is the current live input frame, the system must determine whether to add another row, or column, or both, to matrix D. It does this by checking all the path cost's previous c elements of the current row t and column j of the matrix. If the lowest cost is in a row, it increments the row. If the lowest path is in the column, it increments the column. If the lowest cost is $D(t, j)$, the current cell, it increments both. Also, if a row or column has been incremented $MaxRunCount$ times, it then increments the other, thus preventing the system from running away in one direction. This implementation sets $c = 500$ and $MaxRunCount = 3$ as described in [5].

Indices t and j are pointers to the current real-time position in U and V. At any point $U(t)$ (the current frame in the real-time input), the value of j is the current estimated location in V. Since index t is the current live input frame, it will always increment steadily. Index j, however, will increment based on where the online time warping algorithm estimates the current temporal location to be in the reference recording. AD is inserted based on the real-time current value of j.

3 The Alignment Process

Three inputs are needed to trigger AD: the reference recording V, one or more frames of ongoing live input U, and an array of frame numbers F, where $F(1...x)$ represents the frame at which AD number x should be triggered. U and V are arrays of feature vectors. Both $U(n)$ and $V(n)$ are a single feature vector at frame n. Prior to the live performance commencing, all features of the reference recording are extracted and placed in V. U is extracted in real time during the live performance.

When the show begins, $t = 1$ and $j = 1$, which are references to the indices of the first frames of U and V, respectively. Each time t increases (meaning the live recording has progressed by one frame), the new value of j is determined, based on the online time warping algorithm. If the algorithm determines that U is progressing faster than V at that moment, then $t+ = 1$. If U is slower than V, then $j+ = 1$. If they are both moving at the same speed at that moment, then both t and j are incremented. Index j will keep increasing until it matches t's estimated location, and a new t (live input frame) is obtained (or, alternately, j will not increase while t catches up). The AD number x is triggered when $j = F(x)$. In this way, the descriptive tracks are able to align with the live performance based on the online time warping's estimation of the current index j of the reference.

Since the actual size of the matrix is unknown, an empty square matrix of 40,000-by-40,000 was created in the current implementation, which holds approximately 25 minutes of data on either t or j given the feature extraction parameters presented earlier. During the alignment, when one index is incremented up to the size of the matrix, the matrix is reset and the path cost returns to 0. In this manner, the alignment can run indefinitely, and the calculated path cost

does not increase indefinitely. During the feature extraction phase, the MFCCs for each minute of audio was calculated and extracted in less than 2 seconds while running on a 2.7 GHz MacBook Pro using a C/C++ implementation. Offline tests of the online algorithm were able to process one hour of alignment (including extraction and matrix calculation) in about 4 minutes. This process therefore runs comfortably in a real-time scenario.

4 Evaluation

To evaluate this method, two different audio recordings of the same theatrical productions were used, with one recording as a reference and the other as live input. Markers were placed manually in both recordings to represent specific moments in the production (such as lines of spoken words or musical cues). The algorithm was then run in real time and the mark locations found during the alignment were compared to the actual mark locations in the live input.

In the first evaluation experiment, two recordings of Gilbert and Sullivan's *H.M.S. Pinafore* were used: a D'Oyly Carte recording from 1949 and a Malcolm Sargent recording from 1957. In both recordings 213 specific points were marked; these points were meant to simulate where AD may be triggered. This experiment used the D'Oyly Carte version as the reference and the Malcolm Sargent version as the live input.

After completing the alignment, the ground truth was compared with the marks automatically located by the algorithm. A total of 161 marks (76%) were found less than 1 second from the mark's actual location in the reference; 183 marks (86%) were found less than 2 seconds from the actual location; and 200 marks (94%) were less than 5 seconds. The mean difference between the marks indicated by the algorithm and the ground truth was 1.2 seconds (SD = 2.68 seconds).

To test the algorithm in a more realistic situation, a second experiment using two recordings with notable, audible differences were obtained: the Broadway (2007) and London (2010) cast recordings of *Legally Blonde, The Musical*. The London version was recorded in front of a live audience and contains audience noise, laughter, ambience, etc. that is not present in the Broadway recording. The only alteration made to the recordings was the removal of one track from the London version because it was out of order in the Broadway recording.

In both versions of *Legally Blonde*, 169 locations were manually marked, and the alignment was run in real time using the London recording as the reference, and the Broadway recording as the live input. The results showed that 117 marks (69%) were found within 1 second of the reference, 133 (79%) were within 2 seconds, and 147 (87%) were within 5 seconds. The mean difference between the generated marks and ground truth was 1.79 seconds (SD = 3.32 seconds).

In both experiments, the total duration of each recording was over an hour, and the algorithm was able to keep up with the long live input, and automatically correct itself after moments of difference between the reference and the live input. If there is a "mistake" between productions and the AD becomes misaligned,

Table 1. Results of the two evaluation experiments. Marks refers to the total number of annotated marks for each show; values in the < 1, 2, 5 sec columns indicate the percentage of marks found within 1, 2, and 5 seconds of the ground truth; St. dev refers to standard deviation of the differences between the found marks and the ground truth.

	Marks	**< 1 sec**	**< 2 sec**	**< 5 sec**	**St. Dev**
Pinafore	213	75.57%	85.92%	93.92%	2.68 sec
Blonde	169	69.23%	78.70%	86.98%	3.32 sec

the algorithm may correct itself as the production progresses. For example, the longest difference between reference and live for all experiments was about 21 seconds, which occurred during a significant period of divergence between the two recordings of *Legally Blonde*. However, the algorithm was back to correctly aligning once again less than 2 minutes later, with the next marks falling 120 ms from the reference.

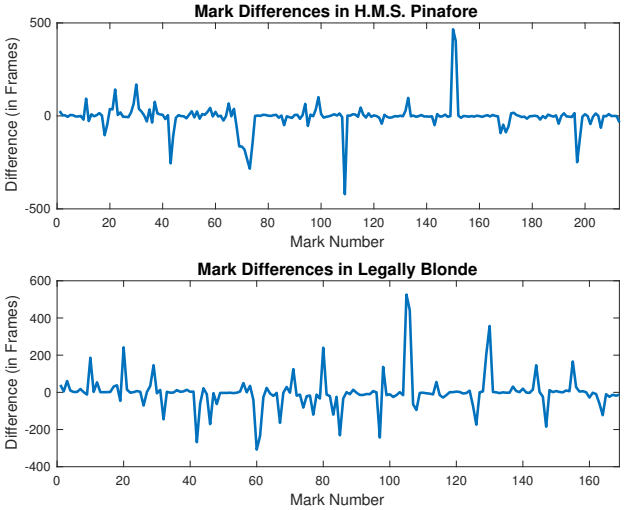


Fig. 1. Accuracy of all marks shown for *H.M.S. Pinafore* and *Legally Blonde*, shown as deviations from ground truth in frames. X axis indicates mark number; Y axis indicates difference in number of frames (25 frames = 1 second).

Within the context of a live theatrical environment, these results show that most (79-86%) AD will be triggered within two seconds of the actual event occurring. These metrics indicate that theatergoers would be able to follow the visual elements of the production in a timely way.

5 Implementation

Software implementations to increase the availability of audio description generally take the form of automating some task of AD creation or deployment, thus decreasing cost and complexity, and ultimately increasing availability. For example, the *CineAD* system [4] uses information from existing closed captions and a teleplay or screenplay of a fixed film or television program to generate a descriptive script automatically, which may then be read by a synthetic voice or by a human. Alternately, *LiveDescribe* [3] seeks to recruit amateur describers to describe videos; it does this in part by analyzing a video and allowing describers to record descriptive audio only during breaks in dialogue.

In a live setting, and in particular when a live performance is imperfectly repeated, automated AD must accommodate for variations, but must still be able to follow some static representation of the performance in order to correctly align the description. Thus, we propose both a software implementation and workflow that takes into account existing theatrical technology (audio recording) as well as established music information retrieval techniques (feature extraction and online time warping) to align AD in real time. The previous sections of this paper have discussed the algorithm and parameters of the software; this section describes how the software may be ideally used in a live setting.

Computer and Theatrical Requirements. The software in this implementation runs comfortably on a 2012-era MacBook Pro with a solid-state hard drive and 16 gigabytes of memory. The computer must be able to obtain two channels of mono input simultaneously—one input to capture the reference recording and (later) the live input, and the other to capture the live audio describer. Importantly, both channels must be isolated from each other such that they do not capture audio from each other, meaning the describer must be listening with headphones so that the reference recording is not captured on the descriptive audio track. The computer must also have a single mono audio output channel, though this will not run simultaneously with the input. This output is transmitted (by either wireless or some other mechanism) to audience members using the descriptive service.

Software Interface. The software's core functions are *Record* and *Begin Show* buttons. The *Record* button activates the two mono inputs simultaneously to capture both the reference recording and the descriptive audio. The software records the reference's alignment to the AD by detecting when the speaker begins to talk, and the corresponding frame number is retained for the particular mark. Alternately, a third function, *Add Mark*, can be manually activated each time the describer wishes to add a new mark.

The software may have supplemental editing features. The ability to replace portions of the reference recording may be helpful if a long-running show has scene changes. Additionally, the ability to minutely correct the trigger timing

of specific descriptive audio may help with fine tuning the AD between performances.

Capturing the Reference and Audio Description. With the software running, the setup configured as above, and the show commencing, the live describer presses *Record* on the computer and the system begins to capture both the performance and describer's voice. The system records the entirety of both mono inputs, capturing the sample numbers for each audio described event. By the end, the system will have captured all relevant data needed to play back the description to future audiences. After this point, the human describer is no longer needed, and the system may be automated.

Providing Audio Description to Audiences. A theatrical technician activates the *Begin Show* button as the production commences. At this point, the mono output (to the wireless receivers of audience members) and the mono input for the live recording (which is the same input as for the reference) is activated. The online time warping process is activated, and the AD is triggered at the correct moment.

6 Conclusion and Future Work

In this paper we presented an automated approach to triggering audio description for a live musical theater performance using audio from a previous performance and an online time warping algorithm. The method is able to correct itself and adapt to variations between the reference and live performance, which is necessary for an effective real-time method. Although the method could be further refined in the future by taking into account very large variations due to intermissions and audience applause that are typical in live performances, the evaluation experiments showed that significant differences such as changes in casting, script, and instrumentation, are already handled robustly.

The software implementation of the system described here can be integrated into a theater's existing setup with minimal interference. For example, the system is able to capture, process, and deploy AD independent from the existing software or hardware controls, since it only uses an audio signal which is readily available in the performance space. Other than an initial reference recording and the descriptive tracks themselves, no other setup or configurations were required. The simplicity of the method's technical setup and its overall flexibility provide a new way to make theater experiences for visually impaired audience members more inclusive and accessible.

Given the proliferation of accessibility on personal computing devices, using a smartphone to align and deliver the description would improve the overall success of the system. Being able to track a live performance from a mobile device without having to be connected to a wireless transmission mechanism would allow AD to be completely in control of the user, not reliant on the setup of the theater.

Audio description is a relatively new but quickly expanding accommodation for those with visual impairments. While it is becoming more common in film and television, AD for live theatrical performances remains rare. Decreasing the cost and complexity of creating and deploying AD would increase its availability, thus making the enjoyment of live theater more accessible to blind and visually impaired individuals.

References

1. Arzt, A.: Flexible and Robust Music Tracking, Ph.D. dissertation, Johannes Kepler University, Linz (2016)
2. Arzt, A., Widmer, G., Dixon, S.: Automatic Page Turning for Musicians via Real-Time Machine Listening. Proceedings of the European Conference on Artificial Intelligence, pp. 241-245 (2008)
3. Branje, C. J., Fels, D. I.: Livedescribe: can amateur describers create high-quality audio description? Journal of Visual Impairment & Blindness. Vol 106, No. 3, pp. 154-165 (2012)
4. Campos, V. P., de Araujo, T, M, U., de Souza Filho G.L., Goncalves, L M. G.: CineAD: a system for automated audio description script generation for the visually impaired. Universal Access in the Information Society, pp. 1-13 (2018)
5. Dixon, S.: Live tracking of musical performances using on-line time warping. Proceedings of the 8th International Conference on Digital Audio Effects, pp. 92-97 (2005)
6. Fryer, L: An Introduction to Audio Description: A Practical Guide. Routledge (2016)
7. Kumar, D. S. P.: A simple MFCC extractor using C++ STL and C++11. Source code at `http://www.github.com/dspavankumar/compute-mfcc` (2016)
8. Lertwongkhanakool, N., Kertkeidkachorn, N., Punyabukkana, P., Suchato, A.: An Automatic Real-time Synchronization of Live Speech with Its Transcription Approach. Engineering Journal, Vol 19, No. 5, pp. 81-99 (2015)
9. Litsyn, E., Pipko, H.: System and method for distribution and synchronized presentation of content. U.S. Patent Application 16/092,775 (filed May 2, 2019)
10. Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling. ISMIR. Vol. 270. (2000)
11. Muda, L., Begam, M., Elamvazuthi, I.: Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing, Vol 2, Issue 3. (March 2010)
12. Plaza, M.: Cost-effectiveness of audio description process: a comparative analysis of outsourcing and "in-house" methods. International Journal of Production Research, pp. 3480-3496 (2017)
13. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimisation for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 26, pp. 43?49 (1978)
14. Snyder, J.: The visual made verbal: A comprehensive training manual and guide to the history and applications of audio description. American Council of the Blind. (2014)
15. Szarkowska, A.: Text-to-speech audio description: towards a wider availability of AD. Journal of Specialised Translation. Volume 15, pp. 142-162 (2011)

Jean-Claude Risset and his interdisciplinary practice: what do (or could) the archives tell us?

Vincent Tiffon[1],

[1] Aix Marseille Univ, CNRS, PRISM, Marseille, France
tiffon@prism.cnrs.fr

Abstract. In 2017, Jean-Claude Risset gave his archives to the PRISM's laboratory. Thereby the researchers' community will have soon at their disposal a fund, especially interdisciplinary art and science oriented. For the moment, the archives are divided into two main parts: one within scientific research and one within artistic creation activity. More specifically, Jean-Claude Risset's own story shaped major interdisciplinary orientations: first of all, his pioneering research at Bell Labs, then back to "french reality" (his half-failure with Ircam and his difficulties concerning Marseille-Luminy), afterwards his quest for solutions as a political lever, especially through the Art-Science-Technology's report in 1998, and finally his turning point with his CNRS 1998 Gold Medal, consequently increasing conferences and mostly concerts. In addition, the study of material aspects (sharing activities between the laboratory and his home, place and content of documentation, etc.) is also necessary to understand "Risset's practice" of interdisciplinary.

Keywords: Musicology, Archives, Science & Art

1 Introduction

First of all, I would like to thank Richard (Kronland-Martinet), and through him the PRISM laboratory, for their welcome when arrived last June. I would also like to warmly thank Nemanja Radivojevic, PhD student from the University of Bern (in Switzerland), who came at the end of July to consult the Risset's fund, and helped me a lot in organizing it. Finally, I would like to thank Tanguy Risset for his trust and his help during this summer.

The physician and composer Jean-Claude Risset gave his archives to the PRISM's laboratory in 2017, in association with the INA-Grm in Paris for extracting and securing the scores, the audio-supports and digital documents (correspondences, patches, etc.). Thereby the researcher's community will have soon at their disposal (when it will be organised) a fund especially interdisciplinary art and science oriented. More precisely, the study of these archives will help all of us to better understand how Risset organised his double activity as a researcher in science and as a composer. It will also make it possible to evaluate in detail the way Risset conceived interdisciplinarity. Incidentally, these archives will also make it possible to work on the analysis of composition processes, in a perspective of TCPM conferences (Tracking the Creative Process in Music). Indeed, these archives could be a kind of workbook to implement interdisciplinary within a team or across team.

After a quantitative presentation of the collection (for the moment relatively approximate because the process is ongoing), we will propose a first qualitative approach, aiming to suggest some hypotheses on Jean-Claude Risset's own choices concerning Art / science interdisciplinarity.

2 Quantitative Content

With about 80% of the archives retrieved, we can describe its content as follows. These figures are a low estimate.

- 39 original manuscripts (other manuscripts are also being processed at l'INA-Grm de Paris). For the record, the number of Jean-Claude Risset's works is 68
- Sketches, drafts and various documentations of 73 works ranging from 1963 to 2016 (including several projects of unfinished parts, at least 5)
- Notebooks (more or less dated), schedules (from 1950 to 2016)
- 1.5 linear meters of archives about scientific topics (for examples, auditory perception, “hot articles” on perception, auditory illusions, historical articles on synthesis, motricity and musical performance, Ircam (1973-1979), quadraphonic spaces, voice, hearing, correlation, filtering, heuristic creativity intelligence, wavelets, math and music, etc.
- 2-4 linear meters of archives about administration of research (reports, evaluations, project reports, laboratory staff administration, careers, etc.)
- 3-5 linear meters of archives of conferences and / or concerts
- 2 linear meters of archives of projects of talks or/and papers, drafts of talks or/and papers and articles
- Between 2 and 4 linear meters of documentation about works performed in concert
- 7 linear meters of drafts, sketches and runs by Music IV, Music V, screenshot of Max-MSP patches (for works with Disklavier/Yamaha)
- 1 linear meter of class notes (when he was student)
- 44 linear meters of books, of which we can make a rough division between
  - a) the fields of Art-science, perception, psychoacoustics, cognitive sciences, computer and music... (22)
  - b) music and musicology (19)
  - c) other scientific fields (3)

    Books not related to the professional activities of Risset remained in the private domain of Jean-Claude Risset's heirs.
- 1.5 linear meters of correspondence
- undetermined number of recordings [not yet stripped] (concert recording, concert master

3 Qualitative Content

3.1 Material Perspective

The question of the combination of scientific and musical activities naturally depends on the places where these activities have been carried out, whether or not they have been correlated. Here again, we will distinguish different periods in Jean-Claude Risset's professional life.

On one hand, the "Bell Labs" (1964-65, 1967-1969) and "Ircam" (1975-1979) periods are times of clearly joint activity, precisely because of the claimed interdisciplinary nature of these two institutions. Resident invitations to the CCRMA (1971, 1975, 1982, 1986, 1998), Dartmouth College or the MIT Media Lab (1987, 1989) belong to the same kind of interdisciplinary. The close combination of theoretical research, technological development, musical "craftsmanship", collaborative work with teams, at the heart of the interdisciplinary approach defended by Anglo-Saxon researchers in those years, had a very strong adhesion in Jean-Claude Risset's mind.

On the other hand, his activity within the LMA (Laboratoire de Mécanique et d'Acoustique, CNRS, Marseille, between 1971 and 1975, then from 1979 to his death), reveals a more strong division between the scientific work, carried out within the laboratory's premises (Luminy, then CNRS-Aiguier campus), and his artistic activity, rather made from home. These first findings, which we will have validated them through a series of interviews with direct witnesses of his Marseille activity, interviews that I intend to conduct in 2020. In a rather obvious and logical way, it can be recalled that the sound synthesis on the Music software is carried out in the CNRS laboratory because it requires heavy equipment before the Personal Computers technological mutation, and the parts of graphic writing for instruments (for mixed music) or the writing of purely instrumental works have been made rather at home. This point has been confirmed by Tanguy Risset (Jean-Claude Risset's son). By "home", we refer to Marseille, or in summer in Bénodet in Britain, West of France.

The observation of the existing documentation at his home - articles, and especially books, excluding books outside the artistic and scientific fields - shows the predominance of musicological books, and the fields of cognitive sciences, computer music and sound perception, many of which have been already there in the LMA laboratory (now PRISM laboratory). A more in-depth study of the home documentary collection moved by the Risset's family to the PRISM lab still to be undertaken.

3.2 Disciplinary Perspective

"Both science and art are ways of understanding the world beyond the individual subjective view of reality" said Jean-Claude Risset [4, 18]. And Risset to precise: "According to Jacques Mandelbrojt [6], science is good at describing reality itself ("en soi"), while art is more apt to express or translate reality in us ("en nous"): its exquisite sensitivity and its extreme quest of subjectivity capture human universals

and archetypes" [4, 18]. For the record, Jacques Mandelbrojt is both painter artist and theoretical physician.

Beyond the common vocation of science and art to create knowledge, the interdisciplinarity according to Risset was first and foremost a radical mastery of each of the two disciplines. "I am a composer and researcher. Inseparably, while never having confused [scientific] research with [artistic] creation" [4, 173] With his difficulties: "If my research has nourished my creation, I have nevertheless experienced the conflict between different activities in their aims, methods and tempo. This conflict is transposed to the institutional level" [1, 175]. These are also areas that are largely different and even often opposed, by their goals (revealing mechanics vs. inventing new sound and musical forms), their methods (theory and experiential vs always *experimental*), their inclusion in the social ecosystem: "The scientific work is collective, ascertainable, provisional or temporary: it is subject to correction (adjustments), obsolescence and incorporation into the progress of science. The artistic work, on the other hand, is individual, subjective and sustainable." [1, 179] "[Science] proceeds by tests superimposed one on the other and whose dark thickening slowly shows at the level of the true. Nothing similar in art. 'Art is not successive, art is once and for all', said Victor Hugo [2, 29], 'real presences' according to George Steiner [3]." This vision, close to the thinking of the physician Jean-Marc Levy-Leblond [5], a vision that no one could suspect of being too positivist, emphasizes the notion of *alloy* rather than that of *identity*.

Risset said: « There is no need to recall that music has benefited from science and technology. However, the inspiration which music has brought to science and technology is largely under-estimated. Music has a special kindship with scientific disciplines. According to Jon Appleton, 'music inspires the kind of rational thoughts necessary to produce scientific work'" (4, 13). However, we suggest that the reading of the draft articles, sketches of theoretical works, as well as some of Risset's texts shows that Jean-Claude Risset participated very largely in fighting against this asymmetry, he was favor of cross-fertilization art and science.

Finally, this dual activity is also expressed in the writing of reports and other evaluations, many of which have been found in the archives. Jean-Claude Risset's extreme generosity meant that, when he was asked to take part on PhD defense or HDR juries (HDR means French new thesis after the PhD thesis to obtain authority to conduct research), he did not hesitate to switch from an expertise in musicology to acoustic physics, from aesthetics to signal studies, and so one. Many texts (papers or talks) - including a few excerpts cited here - show a constant self-assessment process, or more precisely, introspection into his practices. He acted as if his own metaposition had to be shared with potential multidisciplinary community suggest us "good practices". Additionally, the very large number of correspondences, scrupulously preserved but totally fragmented in his various activities, shows once again the crosswise dimension of his scientific, artistic, intellectual, and even political relations (but more rarely). A careful study of his correspondence could again allow us to understand some of the mechanisms of interdisciplinary activity.

3.3 Historical Perspective

Jean-Claude Risset's own story shaped major interdisciplinary orientations: first of all, his pioneering research at Bell Labs, then back to “french reality”, afterwards his quest for solutions as a political or institutional lever, and eventually his turning point with his CNRS 1998 Gold Medal, consequently increasing conferences and mostly concerts. As Risset himself admitted, the Bell Laboratories offered him perhaps the only moment of interdisciplinarity to his own person. “I have myself worked as a researcher (at CNRS), as a composer (at IRCAM) and as both (at Bell Laboratories).” [4, 22]. And indeed, according to Jean-Claude Risset, Bell Labs seemed to be a model: “It must be said that the hosting of American laboratories, not only the material capacity, but also the open-minded and interdisciplinary affordability, were quite large and, in my opinion, quite exemplary, and this is not the case here [in France]. That is to say, in this great extraordinary scientific laboratory at the time called Bell Laboratories, we were involved in information theory, that is, we discovered the background noise of the Big Bang, we discovered the theory and practice of the transistor, we would never stop recalling the extraordinary contributions. There was an extraordinary atmosphere of openness where mathematicians, articles, psychologists, physicists, computer scientists, could work together. And so I was able to do both research for music there, and then even music where I did as an artist in residence in a laboratory. In France, it is a difficult concept..."[1].

In fact, second period, is a return to the “French reality”. We must remember his half failure at Ircam (1975-1979). The requested and insistent injunction of "artistic production" was incompatible, according to Risset, with the longtime of scientific research. What do the archives say? It was a time of great musical production with *Inharmonique, Passages* and works around the emblematic *Songes* (*Profils*. *Moments Newtoniens*, etc.) Hence Risset's remark mentioned above that his Ircam’s activity was essentially related to musical composition.

But the return to France was also the initial enthusiasm for the creation of an interdisciplinary department (art-science) within the Science Faculty of Marseille-Luminy, under the impetus of the physicist Daniel Kastler (son of the Nobel Kastler Prize winner) and the support of Mohammed Mebkhout, for 4 years (between 1970 and 1975). The Ircam parenthesis (1975-1980) being closed, his scientific activities were only within the LMA. The archives show both the richness of the correspondence with these Marseilles precursors of the introduction of interdisciplinarity in France, but also the administrative difficulties with which Risset faced. The lack of understanding of the CNRS's supervision at the time, which only recognized research results, not musical activity, hindered cross-fertilization. And yet, "It is for musical reasons that our computer music team [IM means “Informatique Musicale” - LMA, CNRS Marseille] has contributed since 1984 to the development of the possibilities of wavelet transformation [7]"[1], to cite just one example.

Afterwards, Risset quest for solutions as a political, institutional and academic education lever, especially through the Master ATIAM (Acoustic, Signal processing and Computer Science Applied to Music) from 1993, and the Art-Science-

1 Interview with Jean-Claude Risset by Jean-Yves Bosseur, France Culture, « Opus », 17/04/1999.

Technology's report in 1998, commissioned by the Minister of “Education Nationale” in France. The archives are particularly rich in files mentioning numerous contacts at national level, for the implementation of the ATIAM Master's degree, then the broad consultation with the main actors of the art-science-technology in France (following the creation in 1983 of an association, "Collectif pour la Recherche en Informatique Musicale").

The turning point with his CNRS 1999 Gold Medal, consequently increased conferences and mostly concerts. Concerning this period, the archives are extremely abundant (numerous invitations, especially outside France, which combine talks or keynotes and concerts (monograph or isolated pieces). Jean-Claude Risset scrupulously kept all the drafts of the talks, the correspondence for the preparation of the concerts, and other more tourist details.

4. Conclusion

"I think that interdisciplinary never works as good as when it is embodied in the same person. There are already communication difficulties within myself between the musician and the scientist, even though I can be both but not necessary at the same time. But then if it’s about getting around a table... for instance, interdisciplinary with specialists if each one of them doesn’t go part of the others way, it seems extremely heavy to me"[2]. Interdisciplinarity is not so frequent, neither at the individual level, nor at the institutions level. Concerning Jean-Claude Risset, the study of archives (for the moment superficially) seems to indicate that Risset composer is more the inspiration of Risset researcher (than the opposite). Thus, Risset's work seems to seek to rebalance frequent tropism in the field of art and science, according to which scientific discoveries inspire artists, rather than the opposite. Therefore, musical ideas can largely provoke new fields of research and knowledge.

5. References

1. Risset, J.C.: Recherches au-dessus de tout soupçon. In: Autrement, n°158 (1995)
2. Hugo, V.: Art et science. Paris, Acte Sud (1985) – issue de William Shakespeare (1864)
3. Steiner, G.: Reel Presences. Is there anything in what we say?. London, Faber and Faber, (1989)
4. Risset, J.C.: Science, Technology and Art as Mutual Inspirations: the Computer as an Interface. In SAT 2006, pp. 13-23 (2006).
5. Levi-Leblond, J.M., La science n’est pas l’art : brèves rencontres. Paris, Hermann (2010)
6. Mandelbrojt, J.: Les cheveux de la réalité – autoportraits de l’art et de la science. In Alliage, Nice (1991).
7. Kronland-Martinet R., Morlet J., Grossmann A.: Analysis of sound patterns through wavelet transforms. In: Int. Journal of Pattern Recognition and Artificial Intelligence, vol. 1, pp.273-302 (1987)

2 Interview with Jean-Claude Risset by Philippe Boulanger, France Culture, « la Science et les hommes » 10/07/1991.

Spatial perception of Risset notches

Julián Villegas[1]

University of Aizu
julian at u-aizu.ac.jp

Abstract. The apparent movement of Risset tones and Risset notches (i.e., the opposite of Risset tones, replacing silence by noise and frequency components by notches) are investigated. Contrary to previous findings, no significant differences between horizontal and vertical movement associations were found, regardless of stimuli. However, whereas the tones were more likely to be subjectively associated with approaching sources, notches were associated with receding ones. The direction of frequency glide also had a significant effect: ascending glides were more likely to be associated with horizontal movements and descending ones with vertical movements. These findings suggest that although both stimuli evoke similar illusions, the perception of their spatial attributes are different.

Keywords: Risset tones, Risset notches, Pratt Effect, Doppler Illusion, Auditory Movement Perception.

1 Introduction

The frequency of a tone is usually associated with a corresponding location in the vertical plane: low frequencies with low heights and vice versa. This association is known as the Pratt effect [1]. When a series of tones ordered by frequency is presented, ascending series seem to be associated with sources traveling from low to high elevations, and descending series seem to be associated with sources traveling in the opposite direction [2].

Associations of frequency and locations (or movements) are not exclusive of the vertical plane. Subjects often associate increasing frequency with sources approaching them in the horizontal plane. This is known as the Doppler illusion [3]. This association is contrary to the physics of the phenomenon: Imagine a bystander next to a road where a vehicle is approaching, the frequency f' observed by a stationary listener located at a distance from the trajectory of a sound source emitting a constant frequency f and approaching in a straight line at a constant subsonic speed s_s is

$$f'(t) = \left(\frac{c}{c - s_s \cos(\theta)}\right) f, \tag{1}$$

where c is the speed of sound, and θ is the angle formed by a ray from the source to the listener and the trajectory of the source at a given time. As illustrated in Eq. 1, f' is always decreasing, and yet it is reported as increasing.

Situations arise where the two associations compete: an ascending sweep-tone may be associated with a source traveling from low to high in the vertical plane by account of the Pratt effect, but the Doppler illusion suggests that the same tone could be associated with an approaching source in the horizontal plane.

Risset tones [4] can be considered as a collection of simultaneous sweep-tones that are separated in frequency by the same interval (usually, one octave). The lack of harmonics in the timbre created by a Risset tone makes the resulting sound ambiguous in pitch height, effectively creating the illusion of an ever ascending/descending pitch according to the direction of frequency changes in the tones.

In previous research [5], [6], Risset tones were found to be more likely to be associated with horizontal movements than vertical ones. Ascending Risset tones were more commonly associated with approaching sound sources, and descending Risset tones with receding ones. These findings suggest that, at least for Risset tones, the Doppler illusion takes precedence over the Pratt effect.

One possible explanation for these associations is pitch dependence on sound level. Increasing the level of low/high frequency tones yields changes on their apparent frequency in the same direction (i.e., low frequency tones are reported lower and high tones higher) [7], [8]. According to Stevens [7], the underlying mechanism for this pitch shift is that at high levels, the point of maximal stimulation in the basilar membrane is shifted towards the basilar nearer end.

Further studies clarified the pitch–intensity association for complex tones and sweeps similar to those in Risset tones: the apparent interval size rating reported by musically untrained subjects presented with harmonic sweep-tones at a constant intensity was larger for ascending glides than for descending ones. The presentation level (high or low intensity) also affected the size ratings accordingly [9].

The intensity level of an approaching source increases following an inverse-square law in the free-field, so it is possible that listeners have learned to associate apparent pitch changes with approaching or receding sources depending on the motion-related intensity changes and spectral content of the sound source. These pitch–motion associations may persist even in the absence of level variations.

Other studies have shown that the pitch association with level is rather weak [10], or exclusive of low-frequency sinusoids [11]. Consequently, associations of approaching sources with ascending pitches (or vice versa) may have a more complex explanation than a purely sensory one, including the influence of other sensory modalities like those found in [12] where simultaneous presentation of pitch glides slowed down the detection of visual slopes in the opposite direction of the glide, cognitive processes, etc.

In this research, frequency change associations with apparent movement of sources is investigated using Risset tones and notches, the latter being signals consisting predominantly of energy in all bands (a pink noise) with notches at frequencies where Risset tone components would be observed. Both stimuli evoke a never-ending rise or fall in pitch depending on the direction of frequency glide. It is hypothesized that if the previously observed associations between

glide direction of Risset tones and apparent motion (approaching and receding for ascending and descending glides, respectively) are also observed in Risset notches, it then indicates that the causes of these associations are not perceptual since there is arguably no broadening of the excitation region in the basilar membrane caused by the notches.

This research is important because it helps to better understand auditory movement perception and its association with frequency. Such understanding may help to improve auditory displays in virtual reality environments.

2 Materials

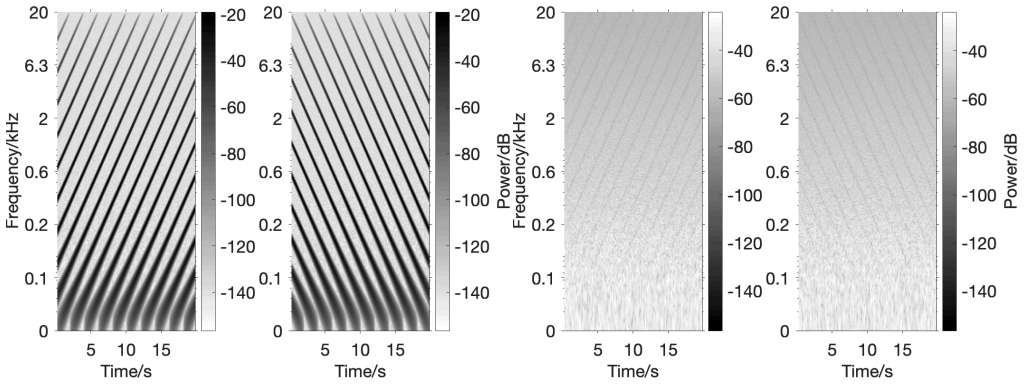


Fig. 1. Risset tones and notches used in the experiment

Participants were subjected to a semi-anechoic monophonic recording of a female speaker reading Harvard sentences [13]. This stimulus was used as a control to verify whether the movement opinions were related to Risset tones or not. Participants were also subjected to Risset tones and Risset notches with the same characteristics of the Risset tones. In all cases, stimuli were presented diotically. Details on the construction of the two latter stimuli are offered in the following subsections.

Risset Tones Ascending Risset tones comprised 20 frequency components separated one octave between them, sampled at 48 kHz, and linearly fade in and out 10 ms. A frequency component x was generated as a logarithmic chirp that traversed the spectrum from $f_l = 20\,\text{Hz}$ to $f_h = 20\,\text{kHz}$ in 20 s:

$$x(t, f) = A(f)\cos(2\pi\phi(t)t), \tag{2}$$

where

$$\phi(t) = f_l \left(\frac{f_h}{f_l} \right)^{\frac{t}{20}} \tag{3}$$

and $A(f)$ is a Hann window:

$$A(f) = \frac{L_p}{2}\left(1 - \cos\left(2\pi \frac{n}{N}\right)\right), 0 \leq n \leq N, \tag{4}$$

where N is the length of the chirp in samples, and the peak level $Lp = 20\,\text{dB}$ is centered at 632 Hz (i.e., between $\text{D}_5^\sharp$ and E_5). Descending Risset tones were obtained as the time-reversed version of the ascending ones. The Risset tones so obtained are presented in Fig. 1a.

Risset Notches Risset notches were created by filtering a pink noise (power spectral density $1/f$) with a series of notch filters as suggested in [14]. Length of the signal, separation between notches, sampling frequency, and fading time, were the same as in the Risset tones.

The pink noise was split into frames of 8192 samples, Hann-windowed, and overlapped $\sim 97\%$ (i.e., a hop size of 256 samples). A notch filter was specified to have a quality factor $Q = 15$. Each frame was notch-filtered at a center frequency f_c determined as in Eq. 3. The notch gain was frequency-dependent in the same manner as in Eq. 4, i.e., the minimum gain ($-25\,\text{dB}$) was obtained at 632 Hz, and the maxima ($-5\,\text{dB}$) at 20 Hz and 20 kHz following a Hann-window shape. Descending Risset notches were time-reversed versions of ascending ones, as before. The resulting signals are presented in Fig. 1b.

3 Method

3.1 Participants

Nine students (including two females) were recruited for this experiment. None of the students had participated in this kind of experiment before. They were 22 years old on average ($SD = 1.80$) and received either financial compensation or credits for a sound and audio course for their participation. Participants had hearing thresholds $\leq 20\,\text{dB}$ (HL) in the range [0.125–8.0] kHz, measured with a Maico MA25 audiometer. These thresholds were considered to be within the normal range of hearing. Permission for performing this experiment was obtained following the University of Aizu ethics guidelines.

3.2 Apparatus

The experiment was conducted simultaneously at five workstations distributed in a quiet room with average $\text{RT}_{60} = 322\,\text{ms}$ and Noise Criterion NC= 25[500]. Participants listened to stimuli via Sennheiser HD 380 pro headphones, calibrated so that a full-scale pink noise yielded a sound pressure level of 69 dB(A).

This level was verified with a Brüel & Kjær 4153 ear simulator and a 2250-Light-G4 sound level meter. Headphones were connected to the audio output of computers running macOS X. The experiment was programmed in a survey tool [15] running on a local server. This survey was presented in a web browser window maximized over a 27″ computer screen.

3.3 Procedure

Sessions started with oral instructions and were about 40 minutes on average (SD = 8.0). Participants were encouraged to ask questions until they considered themselves to be sufficiently prepared for the experiment. Then, they were seated at workstations and asked to follow further instructions presented on the computer screen. Participants were asked to refrain from adjusting their workstation sound level and to carefully listen to each stimulus as many times as they needed to be confident of their opinions. They were also instructed to sit straight and steady, and encouraged to close their eyes while listening to the stimuli. With the exception of adjusting the sound level, we did not enforce any of these conditions.

Overall progress of the experiment was shown as a percentage in a progress bar. The experiment started with a single practice trial (not used in subsequent analyses) followed by the main block. The practice trial was a 10.0 s helicopter sound spatialized as to be approaching from the front via Head-Related Impulse Response (HRIR) convolution, using the generic database described in [16].

The task for the participants had six parts: (1) To rate their agreement with the statement "The sound is outside your head" on a 5-point Likert scale. The extremes of such a scale were labeled as '1: Completely disagree' and '5: Completely agree;' (2) to state whether the sound was 'Stationary,' 'Approaching,' or 'Receding;' (3) to state the apparent origin and (4) destination of a stimulus (in both cases the choices were 'in Front,' 'Above,' 'Behind,' 'Below,' or 'Other'); (5) to state its apparent trajectory (in a straight 'Line,' 'Arc,' or 'Other' path); and (6) to offer comments per trial.

Participants were informed that 'Front' and 'Back' were defined as opposite sides of the coronal plane, and 'Above' and ''Below' as opposite sides of a horizontal plane at the height of the ears. They were also instructed that 'approaching' implied that the final distance to the sound source was shorter than the initial one, and that 'receding' implied the opposite.

Given these descriptions, opinions such as "approaching from above to front in an arc" are ambiguous since they could mean a source approaching from afar and above to a location closer to the face of the listener, or a source high at the zenith moving counterclockwise in the mid-sagittal plane to a location closer in front of the listener. Hence, participants were asked to clarify with comments when their opinions were ambiguous.

Participants indicated compulsory alternatives by selecting options from five drop-down menus, and voluntary comments were typed into text-boxes. Playback started when the participant pressed a play button in each trial. Participants

were free to rewind, fast-forward, repeat, or pause. each stimulus. They submitted their answers by pressing a 'Next' button, which also triggered the start of a new trial. Five stimuli (speech, two Risset tones, and two Risset notches varying on glide direction) repeated five times (25 trials in total) were randomly permuted per participant. No feedback was provided in any case.

4 Results

4.1 Externalization

Externalization ratings were analyzed with a cumulative link mixed model fitted with Laplace approximation. The model assumed that distances to the center of the scale were symmetric. This analysis was eased with the 'ordinal' library [17] in R [18]. Ratings were used as dependent variable, Type (with levels Tone, Notch, and Speech) as fixed factor, and Participant as random factor.

Results of this analysis indicated that Type affected externalization of the stimuli $\chi^2(3) = 51.4, p < .001$. On average, participants were more likely to rate the stimuli as heard outside the head for notches (mean $m = 3.53$) than for tones ($m = 2.89$) and speech ($m = 2.24$).

4.2 Movement

Table 1. Combinations of reported motion and origin/destination used as categories for apparent movement analysis

Plane	Category	Combination
Horizontal	Come	Approaching from the front
	Catch up	Approaching from behind
	Leave	Receding to the front
	Fade	Receding to behind
Vertical	Fall	Approaching from above
	Rise	Approaching from below
	Lift	Receding to above
	Sink	Receding to below

Opinions were manually grouped into apparent movement categories based on compulsory responses and optional comments given by the participants. Apparent movement plane, categories and their meaning are presented in Table 1. Additionally, 'Stationary' and 'Other' categories were included to collect respective opinions. Frequencies of each of these categories per stimulus are presented in Table 2.

Table 2. Opinions grouped in apparent movement categories as presented in Table 1

		Horizontal				Vertical				Stationary	Other
		Come	Catch	Fade	Leave	Fall	Rise	Lift	Sink		
Notch	Asc.	6	1	4	8	0	0	12	3	6	5
	Desc.	2	3	0	5	7	1	3	7	9	8
Tone	Asc.	10	4	0	1	0	4	7	0	15	4
	Desc.	4	2	1	3	13	1	3	6	9	3
speech		1	2	0	0	1	1	1	0	39	0

The statistical analysis of apparent movement was performed with a generalized linear mixed-effects model using a Markov chain Monte Carlo sampler, as implemented in the MCMCglmm [19] library in R.

As expected in the case of speech, 'Stationary' was indicated in 86.67% of the opinions. This is a large proportion in comparison to those of notches (16.67%) and tones (26.67%). Thus, this stimulus was excluded from subsequent movement analysis, limiting it to Risset tones and notches only. Participant was considered a random factor, Type (with levels Tone, and Notch), and Glide (Ascending and Descending, as levels) were considered fixed factors.

As illustrated in Table 2, vertical opinions were more common than horizontal ones (37.22% vs. 30.00%), regardless of stimulus type. However, the 95% highest posterior density confidence intervals for the Type overlaps with zero, indicating no significant difference between opinions of tones and notches ($p = .450$).

A similar analysis showed that regardless of stimulus type, ascending glides were more often associated with horizontal movements (37.78% vs. 28.89%) and descending glides with vertical ones (45.56% vs. 22.22%). This difference was significative ($p = .015$).

Significant differences were also found for apparent motion ($p = .002$), i.e., whether the stimuli was approaching, receding, or otherwise. Notches were more likely to be associated with receding sources than approaching ones (46.67% vs. 22.22%), while this association was reversed for tones (23.33% vs. 42.22%). Note, however, that the proportion of 'Stationary' opinions was larger than the 'Receding' ones for Risset tones. The effect of glide on apparent motion was not significant ($p = .145$).

Finally, no significant effect of Type on the individual categories presented in Table 1 was found ($p = .443$). The effect of Glide on the same categories was also not significant ($p = .830$).

5 Discussion

In previous studies [5], [6], Risset tones were associated with movements in the horizontal plane: ascending/descending glides were associated to approaching/receding sources. In this study, no significant differences between horizontal

and vertical associations were found, but there was a tendency towards vertical associations. Although the actual reasons for this change are still unknown, this finding may be explained by the spectral envelope (i.e., the Hann window centered at 632 Hz) used in this experiment. This envelope was originally devised by Shepard [20] to give a strong sense of pitch at the center frequency of the filter, but it is ultimately unnecessary to evoke an ever-rising/falling pitch, as demonstrated in other studies [21], [22]. In such cases, sensitivity differences along audible frequencies act as a bandpass filter centered at 2–5 kHz. By using this filter though, the glides of the frequency components are less conspicuous and, probably, the apparent movement of the source became ambiguous. On the latter point, note that in contrast with previous findings, in this experiment there was a large proportion of tone opinions in the 'Stationary' category.

It was hypothesized that if associations of tones and notches were similar, the underlying causes of these associations would be not perceptual. On the one hand, their association with movements in the vertical and horizontal planes are similar: ascending glides with horizontal movements and descending ones with vertical movements; on the other hand, tones and notches tend to evoke approaching and receding sources, respectively. Hence, it is not possible to accept the proposed hypothesis, and further studies are needed to elucidate the underlying causes of the observed associations.

6 Conclusions

No evidence of apparent movement association with speech was found. The spectral envelope used in this experiment seems to yield the apparent plane of movement ambiguous. Tones and notches successfully produce the illusion of ever-rising or -falling pitch depending on their glide direction. However, the movement opinions differ across these stimuli, so the perception of their spatial attributes is different. The results of this experiment are insufficient to determine the underlying causes of the associations between changes in pitch and apparent movement of sound sources.

Acknowledgments. Thanks to Naoki Fukasawa for helping with the administration of the experiment. The author also express his gratitude to Prof. M. Cohen for his invaluable comments and suggestions.

References

1. C. C. Pratt, "The spatial character of high and low tones," *J. of Experimental Psychology*, vol. 13, no. 3, pp. 278–285, 1930.
2. O. C. Trimble, "Localization of sound in the anterior-posterior and vertical dimensions of "auditory" space," *British J. of Psychology*, vol. 24, no. 3, pp. 320–334, 1934.
3. J. G. Neuhoff and M. K. McBeath, "Overcoming naïve mental models in explaining the Doppler shift: An illusion creates confusion," *American J. of Physics*, vol. 65, no. 7, pp. 618–621, 1997.

4. J.-C. Risset, "Pitch control and pitch paradoxes demonstrated with computer-synthesized sounds," *J. Acoustical Soc. America*, vol. 46, no. 1A, pp. 88–88, 1969.
5. J. Villegas and N. Fukasawa, "Doppler illusion prevails over Pratt effect in Risset tones," *Perception*, vol. 47, no. 12, pp. 1179–1195, 2018. DOI: 10.1177/0301006618807338.
6. J. Villegas, "Movement perception of Risset tones presented diotically," *Acoustical Science and Technology*, 2019. (in Press).
7. S. S. Stevens, "The relation of pitch to intensity," *J. Acoustical Soc. America*, vol. 6, no. 3, pp. 150–154, 1935.
8. E. Terhardt, "Influence of intensity on the pitch of complex tones," *Acta Acustica united with Acustica*, vol. 33, no. 5, pp. 344–348, 1975.
9. W. F. Thompson, V. Peter, K. N. Olsen, and C. J. Stevens, "The effect of intensity on relative pitch," *The Quarterly J. of Experimental Psychology*, vol. 65, no. 10, pp. 2054–2072, 2012.
10. H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin: Springer, 3rd ed., 2006.
11. Y. Zheng and R. Brette, "On the relation between pitch and level," *Hearing research*, vol. 348, pp. 63–69, 2017.
12. S. Parrott, E. Guzman-Martinez, L. Ortega, M. Grabowecky, M. D. Huntington, and S. Suzuki, "Direction of auditory pitch-change influences visual search for slope from graphs," *Perception*, vol. 44, no. 7, pp. 764–778, 2015.
13. Odeon A/S, "Odeon website," 2018. Retrieved 18 Apr., 2019. Available from `https://odeon.dk/downloads/anechoic-recordings/`.
14. F. Esqueda, V. Välimäki, and J. Parker, "Barberpole phasing and flanging illusions," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)(cit. on p. 111)*, 2015.
15. C. Schmitz, *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany, 2016.
16. T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-Dependent Head-Related Transfer Functions Measured With High Spatial Resolution Using a Spark Gap," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 17, no. 6, pp. 1124–1132, 2009.
17. R. H. B. Christensen, "ordinal—Regression Models for Ordinal Data," 2019. R package version 2019.3-9.
18. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. Version 3.5.3. Retrieved on May 14, 2019. Available from `http://www.R-project.org/`.
19. J. D. Hadfield, "Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package," *J. of Statistical Software*, vol. 33, no. 2, pp. 1–22, 2010. v. 2.29.
20. R. Shepard, "Circularity in Judgments of Relative Pitch," *J. Acoustical Soc. America*, vol. 36, pp. 2345–2353, 1964.
21. S. D. Yadegari, "Self-similar synthesis on the border between sound and music," Master's thesis, Massachusetts Institute of Technology, 1992.
22. M. R. Schroeder, *Number Theory in Science and Communication: With Applications in Cryptography, Physics, Digital Information, Computing, and Self-Similarity*. Berlin: Springer, 5th ed., 2009.

Machine Learning for Computer Music Multidisciplinary Research: A Practical Case Study

Hugo Scurto*[1] & Axel Chemla–Romeu-Santos*[2,1]

[1] STMS IRCAM–CNRS–Sorbonne Université
[2] Laboratorio d'Informatica Musicale, Università degli Studi di Milano
{scurto,chemla}@ircam.fr

Abstract. This paper presents a multidisciplinary case study of practice with machine learning for computer music. It builds on the scientific study of two machine learning models respectively developed for data-driven sound synthesis and interactive exploration. It details how the learning capabilities of the two models were leveraged to design and implement a musical instrument focused on embodied musical interaction. It then describes how this instrument was employed and applied to the composition and performance of *ægo*, an improvisational piece with interactive sound and image for one performer. We discuss the outputs of our research and creation process, and build on this to expose our personal insights and reflections on the multidisciplinary opportunities framed by machine learning for computer music.

Keywords: Multidisciplinarity, Machine Learning, Interface Design, Composition, Performance

1 Introduction

Machine learning is a field of computer science that studies statistical models able to automatically extract information from data. The statistical learning abilities of the models induced a paradigm shift in computer science, which reconsidered mechanistic, rule-based models, to include probabilistic, data-driven models. Recent applications of machine learning led to critical advances in disciplinary fields as diverse as robotics, biology, or human-computer interaction. It also contributed to new societal representations of computers through the loosely-defined notion of Artificial Intelligence (AI).

Computer music also witnessed an increased interest in machine learning. Research has mostly been scientific in focus, using and studying models to automatically analyse musical data—*e.g.,* extracting symbolic information related to pitch or timbre from audio data. This led to technical advances in the field of music information retrieval [1], while also benefiting the field of musicology, notably through large-scale computational analysis [2]. In parallel, machine learning also

* Equal contribution.

enabled the building of many automatic music generation systems, which are currently being invested by the industry in the wave of AI [3].

Importantly, these scientific investigations of machine learning have also enabled the birth of new musical practices. For example, gesture modelling, as a scientific challenge, opened new design perspectives on body-based musical instruments that adapts to one's way of playing it [4]. Similarly, symbolic sequence modelling created new human-machine improvisational situations where the machine learns to imitate a musician's style [5]. Reciprocally, artistic investigations of machine learning began taking a complementary approach, using the models themselves as material for composition of sound [6] and image [7].

We are interested in adopting a *joint scientific and musical approach* to machine learning research. We are inspired by the computer music pioneer Jean-Claude Risset [8], whose research and creation approach to computer science enabled new scientific understandings of sound as a physical and perceptual phenomenon, jointly with an artistic commitment toward the computed aesthetics. His work and personal approach gave insight to both scientists—ranging from formal to social science—, and artists—ranging from composers and performers to instrument designers. Our wish is to perpetuate his multidisciplinary impetus toward contemporary computer music issues related to machine learning.

The work that we present here is a step toward this direction. We led a *scientific* investigation of two machine learning models that jointly frame new data-driven approaches to sound synthesis. We then adopted a *musical* approach toward these models, leveraging their interactive learning abilities to design a musical instrument, which we employed to create an improvisational piece. Rather than seeking general abstractions or universal concepts, our wish was to test these models through a practical case study to engage a personal reflection on the musical representations and behaviors that they may encode. Our hope is that our idiosyncratic research and creation process will help open multidisciplinary perspectives on machine learning for computer music.

The paper is structured as follows. We start by the scientific foundations of our work, describing the two models that we developed for two musical issues—sound analysis-synthesis, and sonic exploration. Next, we present the design of our musical instrument, by describing its workflow and implementation with a focus on embodied musical interaction. We then describe *ægo*, an improvisational piece with interactive sound and image for one performer, which we wrote for our instrument. Finally, we discuss our research and creation process to draw conceptual insight on machine learning for computer music from crossed science, design, and art perspectives.

2 Scientific Modelling

In this section, we describe our two machine learning models, based on *unsupervised learning* and *reinforcement learning*, from a computer science perspective. We explain how they respectively address two specific musical issues: sound synthesis-analysis and sonic exploration.

2.1 Unsupervised Learning for Sound Analysis and Synthesis

Musical Issue. Most sound analysis-synthesis techniques, such as the phase vocoder [9] or the wavelet transform [10], are based on invertible transforms that are independent of the analyzed sounds. Such transforms provide frameworks that can be applied regardless to the nature of the signal, but in return impose a determined structure such that the extracted features are not corpus-dependant. Conversely, could we think about a method retrieving continuous parameters from a given set of sounds, but rather aiming to recover its underlying structure?

Model. The recent rise of *unsupervised generative models* can provide a new approach to sound analysis-synthesis, by considering each item of a given audio dataset $\{\mathbf{x}_n\}_{n\in 1\dots D}$, in our case a collection of spectral frames, as draws from an underlying probability distribution $p(\mathbf{x})$ that we aim to recover. The introduction of latent variables $\mathbf{z}$ allows us to control a *synthesis* process by modelling the joint distribution $p(\mathbf{x},\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, such that these variable act as parameters for the generative process $p(\mathbf{x}|\mathbf{z})$. The full inference process, that would here correspond to the *analysis* part, leverages the Bayes' rule $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$ to recover the distribution $p(\mathbf{z}|\mathbf{x})$, called the posterior.

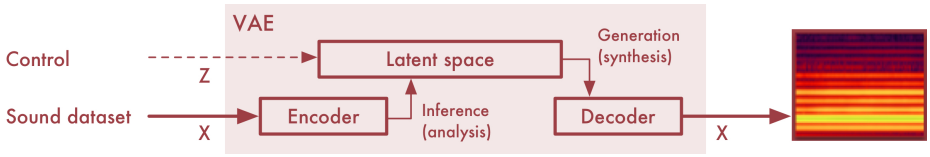


Fig. 1. Unsupervised learning for sound analysis and synthesis. The variational auto-encoder (VAE) encodes a sound dataset into a high-dimensional latent space, which can be parametrically controlled to synthesize new sounds through a decoder.

To improve expressivity of inference and generation, we propose to investigate variational learning, a framework approximating the true posterior $p(\mathbf{z}|\mathbf{x})$ by a distribution $q(\mathbf{z}|\mathbf{x})$, such that both inference and generative process can be freely and separately designed, with arbitrary complexity. The variational auto-encoder (VAE) is representative of such methods [11]. In this model (Fig. 1), inference and generation processes are held by two jointly trained separated networks, respectively the *encoder* and the *decoder*, each modelling respectively the distributions $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$. The inherent Bayesian nature of variational learning forces the smoothness of the *latent space*, a high-dimensional, non-linear sonic space, whose parametric dimensions can be freely explored in the manner of a synthesizer.

In related work, we show how this latent space can be regularized according to different criteria, such as enforcing perceptual constraints related to timbre [12]. We refer the reader to the latter paper for technical details on the model and quantitative evaluation on standard sound spectrum datasets.

2.2 Reinforcement Learning for Sonic Exploration

Musical Issue. Sonic exploration is a central task in music creation [13]. Specifically, exploration of digital sound synthesis consists in taking multiple steps and iterative actions through a large number of technical parameters to move from an initial idea to a final outcome. Yet, the mutually-dependent technical functions of parameters, as well as the exponential number of combinations, often hinder interaction with the underlying sound space. Could we imagine a tool that would help musicians explore high-dimensional parameter spaces?

Model. We propose to investigate *reinforcement learning* to support exploration of large sound synthesis spaces. Reinforcement learning defines a statistical framework for the interaction between a learning agent and its environment [14]. The agent can learn how to act in its environment by iteratively receiving some representation of the environment's state S, taking an action A on it, and receiving a numerical reward R. The agent's goal, roughly speaking, is to maximize the cumulative amount of reward that it will receive from its environment.

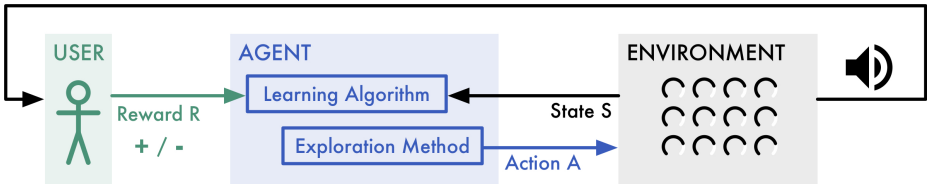


Fig. 2. Reinforcement learning for sonic exploration. The agent learns which actions to take on a sound synthesis environment based on reward given by the musician. The agent implements an exploration method to foster discovery along interaction.

For our case of sonic exploration, we propose that the musician would listen to the agent exploring the space, and teach it how to explore by giving reward data (Fig. 2). Formally, the environment's state is constituted by the numerical values of all synthesis parameters. The agent's actions are to move one of the parameters up or down at constant frequency. Finally, the musician communicates *positive or negative reward* to the agent as a subjective feedback to agent actions. We implemented a deep reinforcement learning model to support learning from human reward signal in high-dimensional parametric spaces [15].

A crucial requirement for reinforcement learning agents is to *autonomously explore their environment*, to keep on discovering which actions would yield the most reward. We developed a statistical method, based on intrinsic motivation, which pushes the agent to "explore what surprises it". The resulting interactive learning workflow was found to be useful to relax musicians' control over all synthesis parameters, while also provoking discoveries by exploring uncharted parts of the sound space. We report the reader to [16, 17] for technical details on the tool and qualitative evaluation from expert sound designers.

3 Instrument Design

In this section, we present our musical instrument that combines our two models and leverages their learning capabilities from a design perspective. We describe how interaction design was framed in joint coordination with hardware and software engineering to support embodied musical interaction.

3.1 Interaction design

Motivation. Our main design motivation was to use our reinforcement learning agent to support musical exploration of high-dimensional latent sound spaces built by our unsupervised learning model.

Specifically, our aim was to exploit the exploration behaviour of our reinforcement learning agent to support non-symbolic *improvisation* inside the spaces. Instead of acting as a tool, we used machine learning as an expressive partner [5] that would be playable by musicians using positive or negative feedback.

A complementary aim was to employ the generative abilities of our unsupervised learning model to support *customization* of sound synthesis spaces. Instead of accurately modelling sounds, we used machine learning as a creative interface [18] that lets musicians experiment with the nonlinearities of the latent spaces.

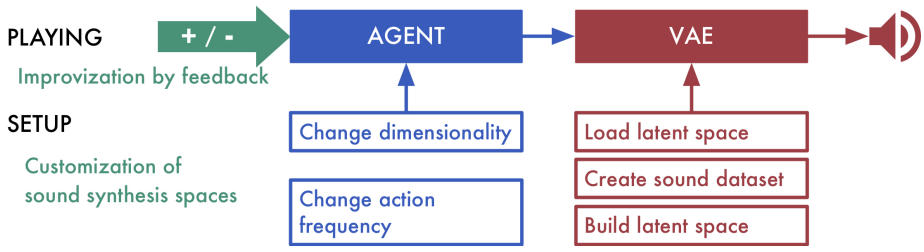


Fig. 3. The interactive workflow that we designed for our instrument.

Workflow. We designed a two-phase interactive workflow, shown in Fig. 3.

The *setup* phase allows musicians to configure the instrument. They can create a customized sound dataset for the unsupervised learning model, experiment with various training parameters, or also load a previously-built latent sound space. They can also change dimensionality of the reinforcement learning agent to explore specific dimensions of the latent sound space, as well as the frequency at which it would take actions inside the latent space.

The *playing* phase allows musicians to improvise with the agent by means of feedback. The agent produces a continuous layer of sound from the spectrum output of the VAE. Musicians can either cooperate with its learning by giving consistent feedback data to attain a sonic goal. Or, they can obstruct its learning by giving inconsistent feedback data to improvise through sonic exploration.

3.2 Engineering

Implementation. Technically (see Fig. 4), the reinforcement learning agent receives a representation of the environment's state S as a position in the latent space $\mathbf{z}$. Then, it takes an action A corresponding to a displacement along some dimension of the latent space. The resulting position has the unsupervised learning model generate a sound spectrum $\mathbf{x}$. Based on the sound, the musician would communicate reward R to the agent. The latter would progressively learn to explore the latent space in relation to the musician's feedback data.

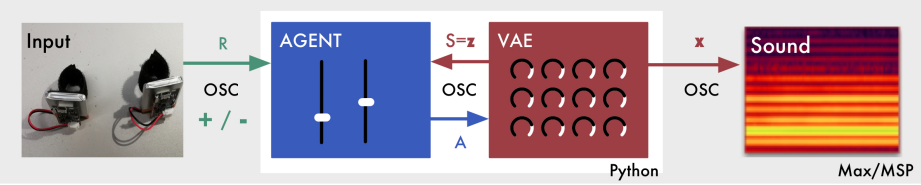


Fig. 4. Schematic representation for the engineering of our instrument.

Hardware. We designed a hardware prototype to support embodied musical interaction (see Fig. 4, left). It consists in two velcro rings, each of them equipped with a wireless inertial measurement unit[1]. We took each unit angular rotation about each forearm axis and summed them to compute a single, normalized numerical reward signal. This, combined with the lightweight, nonintrusive velcro rings, lets musicians experiment with a wide range of gesture vocabulary [19] to communicate positive or negative feedback to the agent.

Software. We implemented our two machine learning models as Python libraries[2,3]. We developed a Max/MSP patch to implement a user interface for the setup phase, as well as a hardware data converter for the playing phase. We leveraged the OSC protocol to bridge hardware data, reinforcement learning agent, unsupervised latent space, and sound spectra together into the patch.

4 Musical Artwork

In this section, we present *ægo*, an improvisational piece that we wrote for our musical instrument, to be premiered at the *14th International Symposium on Computer Music Multidisciplinary Research*, held in Marseille, France. We describe the intended aesthetics of sound, image and body, and detail how composition and performance were approached in relation to our learning instrument.

[1] `http://ismm.ircam.fr/riot/`
[2] `https://github.com/acids-ircam/variational-timbre`
[3] `https://github.com/Ircam-RnD/coexplorer`

4.1 Aesthetics

Motivation. Our artistic motivation for *ægo* was to open a sensitive reflection on what may actually be learned on a musical level through interaction with machine learning, both by the human and its artificial alter ego—the machine. To share this reflection with members of an audience, we opted for a performance format that displays a human and a machine mutually learning to interact with each other—on an embodied level for the human, and on a computational level for the machine—through live improvization.

The learning machine possesses a distinctive musical behaviour, as well as two latent sound spaces, that are all originally unknown to the human performer. The latter will expressively negotiate control of these spaces with the machine, communicating positive or negative feedback using our instrument and its motion sensors placed in both hands. The slowly-evolving spectromorphologies, synthesized and projected in real-time on stage, create a contemplative, minimalist atmosphere intended to let members of the audience freely consider potential learnings of musical qualities by the human and the machine.

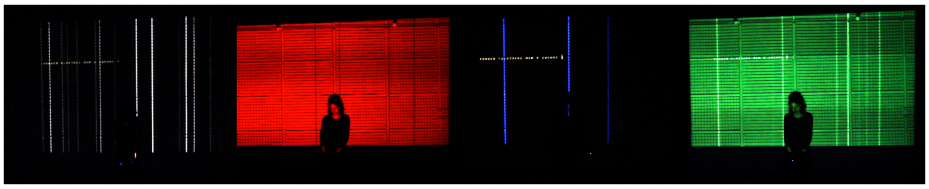


Fig. 5. Pictures taken from *ægo*.

Intentions. The piece's aesthetic intentions toward machine learning lie at three intertwined levels: sound, image, and body (see Fig. 5).

One of our intentions was to reveal the *sound representations* learned by the unsupervised learning model to the audience. We thus built latent sound spaces using sound data that was commonly used and produced in pioneering works of computer music. In addition, we projected the generated sound spectrums on stage to provide the audience with a visual representation that accentuate, not disrupt, the sonic perception of the piece.

Another intention was to display the *exploration behaviour* of the reinforcement learning agent in front of the audience. To do this, we wanted to challenge the skills and abilities usually at stake in performance, by summoning an ecological approach and evoking a sense of reciprocal interaction between the human and the machine. In this sense, rather than using it for control purposes, we used the body of the performer to convey kinesthetic information about how machine exploration may be internally experienced by a human. In parallel, we added raw textual information about the machine's internal state at top left of the image projection to emphasize the machine's encoded perception of the performer.

4.2 Writing

Composition. The piece was composed at three temporal scales (see Fig. 6).

The first scale is that of *exploration*. It consists in the improvisational paths taken by the reinforcement learning agent following the performer's feedback data. We set the frequency of agent actions between 30 and 100 milliseconds. This choice allowed for slow, continuous evolution of spectromorphologies, which enables to grasp the behaviour of the agent inside the latent spaces.

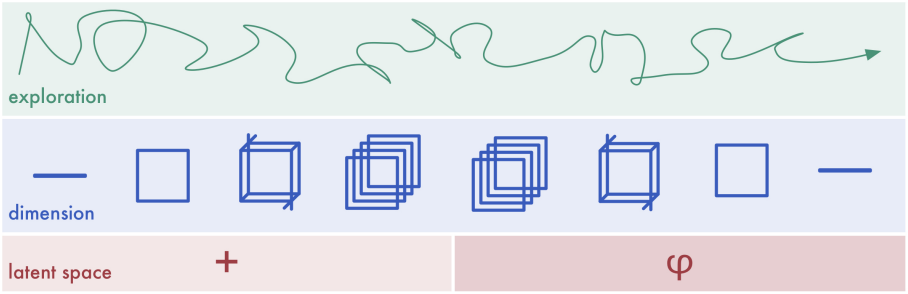


Fig. 6. Temporal structure composed for the piece.

The second scale is that of latent space *dimensionality*. It consists in defining the axis of the latent spaces that the reinforcement learning agent will explore. We set the dimensions to 1, 2, 4, and 8, respectively. This allows to write a specific kind of musical form inside the latent space: the more dimensions we open to the agent, the more sonic variance the performer and audience members will experience.

The third scale is that of latent space itself. It consists in connecting the reinforcement learning agent to another type of latent space. We used two latent spaces, respectively built from additive synthesis sounds and physical instruments recordings (flute, saxophone, piano, violin, bassoon [20]). This enables to write form within different soundscapes, allowing the building of a narrative (here, going from elementary sinusoidal spectra to richer instrumental timbres).

Performance. While the piece is intended to be improvised, our sole instruction toward the stage performer is that he or she globally performs with the machine with an overall sense of attentiveness[4]. We propose that the performer would start the piece facing the audience, relaxed, using the instrument with small forearm rotations only. As the piece would unfold over time, the performer would be free to adapt its gestures in response to the slowly evolving complexity of the explored spaces, focusing on embodied interaction with the machine.

A second contributor is required to manage the two remaining temporal scales of the piece—*i.e.,* changing dimensionalities, and switching latent spaces.

[4]See the following video recording: `https://youtu.be/gCzOoNChlJQ`

5 Discussion

In this section, we take a critical look at the output of our case study by discussing our research and creation process. We then expose our personal reflections emerging from practice with machine learning, and propose conceptual insight for future multidisciplinary inquiries in the realm of computer music.

5.1 Case study

Process. The work presented here relates a practical case study with machine learning in the frame of computer music. We leveraged both conceptual and technical aspects of machine learning to jointly produce *scientific knowledge* with our two models for sound synthesis, as well as *musical creations* through the design of our instrument and the writing of our musical piece. In this sense, our work emerged from a research and creation process, in which we closely articulated a research methodology with a creation project.

We followed a sequential disciplinary agenda (see Fig. 7, solid lines and arrows). We started by the scientific modelling of sonic exploration and sound synthesis, which took us two years to date. We then planned a one-month period to conceive the instrument, write and practice the musical piece. This research and creation agenda was mainly required by our work occupation focusing on computer science research without necessarily addressing music creation.

While many researchers of our laboratory were involved in scientific modelling, we (the two coauthors) managed instrument design and musical piece as a pair. Importantly, we both followed a dual training in science and music, and were doctoral students in the domain of machine learning applied to computer music at the time of writing. In addition, both of us have professional experience in music composition and performance. These dual skills were central to individually work, as well as to effectively collaborate, on conceptual and technical aspects related to machine learning throughout the process.

Output. The relatively short period dedicated to musical creation pushed us to take pragmatic decisions about the form of outputs, notably by relinquishing certain technical developments. For example, using the unsupervised learning model to learn temporal features of sound spectrums could have improved the timbre richness of the generated sounds, as well as supported other musical forms than slow spectromorphology evolution. Also, other agent commands than feedback data could have been designed to support expressive human control over the reinforcement learning agent exploration. Finally, many other musical forms could have been conceived, using other sound datasets—*e.g.,* voice corpora or environmental sounds—and investigating other temporal writings for dimensionality and exploration. Future continuation of our work may consider addressing these research questions to evolve the generated outputs.

5.2 Authors' reflections on machine learning for computer music

Conceptual insight. Beyond the created outputs, our process of practice with the two machine learning models let us reflect on conceptual issues, which feed back into many different disciplines (see Fig. 7, dashed lines and arrows).

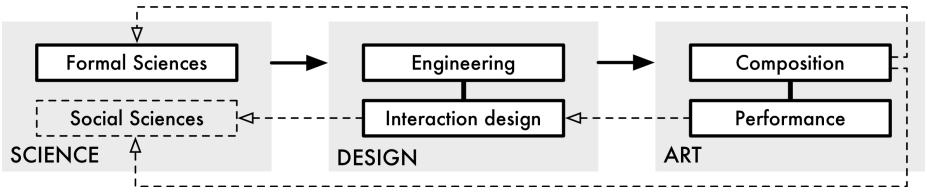


Fig. 7. Our case study. Solid arrows: The sequential research and creation process that we took to scientifically investigate our models, and musically create our instrument and artwork. Dashed arrows: The personal conceptual insight gathered along our process.

On the one hand, composing with the sonic aesthetics produced by the unsupervised learning model let us reflect on epistemological issues that span both formal and social science (Fig. 7, upper and lower dashed arrows). Should machine learning be considered as a modelling tool for sound data, or rather as a framework for sound synthesis that remains to be crafted? Our insight leans toward the latter option. Rather than imposing deterministic rules to define a sound space [21], probabilistic methods propose heuristics that aim to inverse this methods by retrieving structure directly from the data. More specifically, Bayesian approaches filtrates the "space of everything possible" to get closer from the data structure, thus providing interesting generalization abilities in addition to structural information, from the point of view of *formal science*. Conversely, adopting an artistic approach to the learned representations also provides an alternative way of evaluating these models, completing existing machine learning-focused evaluations methods of such unsupervised learning systems. However, such evaluations have to deal with musicological approaches in the realm of the *social sciences*, and remains still an underrated field of research.

On the other hand, performing with a reinforcement learning-based musical instrument offers new design and scientific views on interactivity (Fig. 7, middle dashed arrows). How should we approach an artificial musical partner that learns to behave from our sole feedback data? Alternatively, should exploration be analysed as an expressive musical behaviour? Our insight is that the data-agnostic framework of machine learning may support the development of new modalities for human-machine interaction, which may originate from the *social sciences*. In the musical domain, machine learning may be used to enhance modes of communication that already exist between musicians. Feedback, for example, is a broad communication channel that concern all types of living or nonliving systems [22]. By *designing interactions* with machine learning that rely on feedback data, we may create more accessible musical partners and in

turn instigate analytical views on these embodied notions—as it has been the case with machine learning-based gesture modelling tools [4]. Exploration, as a performative and improvisational practice, remains to be investigated more deeply in that sense.

Toward intrinsic approaches. In this paper our approach was to study the artistic possibilities emerging from the encounter of our two models, rather than to evaluate them separately on their respective tasks. Precisely, our experience in practicing such models revealed to us two distinctive approaches: an *extrinsic* approach, where machine learning models are designed towards a specific task and used faithfully to this end—such as in music information retrieval—, and an *intrinsic* approach, where these models are exploited for themselves and taken as objects that can be explored, hacked, and manipulated—such as in gesture modelling, or improvisational systems. While the first approach has so far been the most common, as machine learning was originally created to tackle complex issues that preceding techniques fell short with, we think that the second may unfold new creative opportunities for computer music, just as Jean-Claude Risset's joint scientific and musical approach to computing did [23]. We hope that the present case study stands in favour of this argument.

While we built on our joint machine learning and music training to lead our case study, it may require more time to manage collaboration between machine learning experts and researchers, engineers, musicians, artists, musicologists, scientists, designers, or epistemologists, toward shared musical goals. We believe that multidisciplinary collaboration is key to lead intrinsic examination of machine learning, and that the latter may be crucial to go beyond suspicions and actively negotiate the place of the human artist in upcoming AI music systems.

6 Conclusion

We presented a practical case study of machine learning for computer music. We studied two machine learning models, from which we designed a musical instrument, and wrote a piece for it. We discussed the research and creation process that fostered our case study and showed the conceptual benefits in terms of feedback. Future work may include multidisciplinary collaborations to intrinsically study machine learning in the realm of computer music.

Acknowledgements

We thank Frédéric Bevilacqua, Philippe Esling, Gérard Assayag, Goffredo Haus, and Bavo Van Kerrebroeck for their broad contributions on the scientific part.

References

1. Hamel, P., Eck, D.: Learning features from music audio with deep belief networks. In 11th International Society for Music Information Retrieval Conference (ISMIR) (2010)

2. Meredith, D. (Ed.): Computational music analysis (Vol. 62). Berlin: Springer (2016)
3. Briot, J-P., Hadjeres, G., and Pachet, F.: Deep learning techniques for music generation-a survey. arXiv preprint arXiv:1709.01620 (2017).
4. Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Gudy, F., & Rasamimanana, N.: Continuous realtime gesture following and recognition. In International gesture workshop, Springer, Berlin, Heidelberg, pp. 73-84 (2009, February).
5. Assayag, G., Bloch, G., Chemilier, M., Cont, A., & Dubnov, S.: Omax brothers: a dynamic topology of agents for improvization learning. Proceedings of the 1st ACM workshop on Audio and music computing multimedia (2006)
6. Ghisi, D. Music across music: towards a corpus-based, interactive computer-aided composition. Doctoral dissertation, Paris 6 (2017)
7. Akten, M., Fiebrink, R., Grierson, M.: Deep Meditations: Controlled navigation of latent space. Goldsmiths University of London (2018).
8. Risset, J.-C.: Fifty Years of Digital Sound for Music. In: Proceedings of the 4th Sound and Music Computing Conference (SMC) (2007)
9. Rodet, Xavier and Depalle, Philippe and Poirot, Gilles : Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions. European Conference on Speech Technology (1987)
10. Kronland-Martinet, R.: The wavelet transform for analysis, synthesis, and processing of speech and music sounds. Computer Music Journal, 12(4), 11-20 (1988)
11. Kingma, D., Welling, M. : Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Esling, P., Chemla–Romeu-Santos, A., Bitton, A. : Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces. DAFx2018 (2018)
13. Ystad, S., Aramaki, M., & Kronland-Martinet, R.: Timbre from Sound Synthesis and High-level Control Perspectives. Springer Nature (2017)
14. Sutton, R. S., & Barto, A. G.: Reinforcement learning: An introduction. MIT press (2018)
15. Warnell, G., Waytowich, N., Lawhern, V., & Stone, P.: Deep TAMER: Interactive agent shaping in high-dimensional state spaces. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018, April).
16. Scurto, H., Bevilacqua, F., & Caramiaux, B.: Perceiving Agent Collaborative Sonic Exploration In Interactive Reinforcement Learning. In: Proceedings of the 15th Sound and Music Computing Conference (SMC) (2018).
17. Scurto, H., Van Kerrebroeck, B., Caramiaux, B., Bevilacqua, F.: Designing Deep Reinforcement Learning for Human Parameter Exploration. arXiv preprint arXiv:1907.00824 (2019)
18. Fiebrink, R., Caramiaux, B., Dean, R., McLean, A.: The machine learning algorithm as creative musical tool. Oxford University Press (2016)
19. Tanaka, A., & Donnarumma, M.: The body as musical instrument. The Oxford Handbook of Music and the Body (2018)
20. Ballet, G., Borghesi, R., Hoffmann, P., & Lvy, F.: Studio online 3.0: An internet "killer application" for remote access to Ircam sounds and processing tools. In: Journées d'Informatique Musicale (JIM) (1999)
21. Chowning, J. M.: The synthesis of complex audio spectra by means of frequency modulation. Journal of the audio engineering society, 21(7), 526-534 (1973).
22. Wiener, N.: Cybernetics or Control and Communication in the Animal and the Machine. MIT press (1965)
23. Risset, J. C., & Wessel, D. L.: Exploration of timbre by analysis and synthesis. In: The psychology of music, Academic Press, pp. 113-169 (1999)

Connecting Circle Maps, Waveshaping, and Phase Modulation via Iterative Phase Functions and Projections

Georg Essl

University of Wisconsin – Milwaukee
essl@uwm.edu

Abstract. In memoriam of Jean-Claude Risset's recent passing, we revisit two of his contributions to sound synthesis, namely waveshaping and feedback modulation synthesis as starting point to develop the connection of a plenthora of oscillatory synthesis methods through iterative phase functions, motivated by the theory of circle maps, which describes any iterated function from the circle to itself. Circle maps have played an important role in developing the theory of dynamical systems with respect to such phenomena as mode-locking, parametric study of stability, and transitions to chaotic regimes. This formulation allows use to bring a wide range of oscillatory methods under one functional description and clarifies their relationship, such as showing that sine circle maps and feedback FM are near-identical synthesis methods.

1 Introduction

Jean-Claude Risset's legacy contains wide-ranging contributions to the field of computer music, sound perception, composition, and sound synthesis. While all these aspects warrant detailed engagement, in this paper we engage only with aspects of his work in sound synthesis. The general range of early contributions of Risset to sound synthesis is staggering and catalogued in his seminal work *An Introductory Catalogue of Computer Synthesized Sounds* [29].

One of Risset's most enduring influences in sound synthesis is seeding ideas of nonlinear functional distortion for sound synthesis (see [29] #150) now known as *waveshaping*. The history of waveshaping is not only intricately tied with an early proposal by Risset, but also with his coming to Marseille. There he would facilitate Daniel Arfib to join, who in turn would develop waveshaping into a full fledged synthesis method [1], an effort that was independently also pushed forward by Le Brun at Stanford University [23]. A number of other contribution helped shape this topic along the way [24, 34, 38]. For an early unifying review of these developments see Roads [31].

Somewhat less widely recognized [42] though equally foundational, Risset also was part of the group of researchers who pioneered the use of feedback and proposed what would later be coined feedback amplitude modulation synthesis [19,29,32]. Risset himself would credit Arthur Layzer, a colleague at Bell Labs for

suggesting the idea (see [29] #510 and #511) though Roads would later attribute the overall development to Layzer, Risset, Matthews, and Moore [32, pp. 244-245]. Since then, feedback has come to play an important role in a range of techniques, such as feedback FM [32, 40] and their variations [21].

Finally, Jean-Claude Risset was undoubtedly fascinated by chaos theory as expressed in his 2014 Keynote at ICMC/SMC in Athens, Greece [30] and the use of fractal and chaos theoretic ideas in his pieces *Phases*, *Strange Attractors* as well as *Pentacle*. Chaos theory can be understood as the most flashy aspect of the study of certain deterministic dynamical systems, of which iterative maps are perhaps the most widely considered form in computer music and sound synthesis [2, 3, 6, 8, 10, 13, 15, 25–28, 33, 37, 41].

The goal of this paper is to develop explicit connections between many established techniques, such as waveshaping, modulation techniques, and a certain class of chaotic oscillators called circle maps [10, 11], hence tying these three strands of Risset's ideas and interests together. Some interrelationships between oscillatory synthesis methods are known, such as the relationships waveshaping [23], phaseshaping [20], frequency [4] and phase modulation and distortion techniques [16, 20, 22]. Some connections between the chaos theory of dynamical systems and classical synthesis methods have already been unearthed. Di Scipio describes his discovery of interesting nonlinear iterative maps and the relation to chaos through the repeated application of waveshaping [36, footnote 16]. It is also known that feedback frequency modulation (FM) [40] can be driven into chaotic regimes [37]. While experimenting with cascaded and feedback FM, Schottstaedt noted that the output will transition into chaos for certain parameter choices [35]. Some of the ideas presented here have already been sketched out but not fully developed earlier [10, 11]. A uniform formulation of all these methods with respect to iterations connects the established sound synthesis literature with the dynamical systems literature and will clarify the individual relationships of different synthesis methods as well as contribute to a broader program to classify synthesis methods with respect to mathematical structure [9].

2 An Iterative Dynamics View of Oscillators on Circles

First we develop a general way to look at oscillatory synthesis algorithms through the lense of dynamical systems. Iterative processes are at the heart of much of dynamical systems theory. But there are a number of further building blocks that add insight into iterative formulations of familiar algorithms. Ultimately we are looking to give the following interpretation of generally familiar phase-centric formulations of oscillators, and interpret them as generalized projections from dynamics on the circle to itself:

$$\underbrace{y_n}_{\text{Time Series}} = \underbrace{p}_{\text{Projection}} \underbrace{(x_n = f(x_{n-1}) \quad \overbrace{\bmod 1}^{\text{Circle Topology}} \quad)}_{\text{Iterative Phase Function}}$$

This will provide a joint view of oscillatory synthesis and the dynamics of iterated functions, which we will explore in more detail as follows.

2.1 Circle Domains as Topological Spaces

Take a circle and any function acting on the circle returning to another position on the circle. It is easy to see that this is a natural, yet general way to think about any sufficiently low-dimensional oscillatory process.

We can write the general form of the circle map as a mapping of the circle $\mathcal{S}^1$ onto itself. It is a general result in topology that quotients of the real line are equivalent to the circle topologically, and this in turn is equivalent to mapping from the repeated interval $[0, 1)$ into itself. Hence, we get three topologically equivalent ways to denote general mappings from the circle onto itself:

$$f : \mathcal{S}^1 \to \mathcal{S}^1 \qquad f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z} \qquad f : [0, 1) \to [0, 1)$$

This final two equations regarding quotients $\mathbb{R}/\mathbb{Z}$ and unit intervals $[0, 1)$ should look already very familiar to anybody who works with sinusoidal oscillators. Equations of the argument of a trigonometric function are periodic with respect to 2π. For example the sine function has the following property $\sin(x) = \sin(x + 2\pi n)$ for all $n \in \mathbb{Z}$. If we divide out the 2π factor we arrive at functions that are periodic on the unit interval $[0, 1)$. This normalization is convenient and will be used through our discussion.

Choosing a topologically closed domain such as a circle provides a domain-induced stability for the data on the domain [10]. All data is bounded to this topology, hence there is no escape, overflow, or explosion possible. This is one reason why, numerically, circle maps are attractive over other chaotic maps, where certain parameter ranges or computational inaccuracies can lead to blowups. Practically, one can induce stability of phase computations by taking the modulo the repetition of the phase. This is a general idea and can be used to stabilize a wide range of synthesis algorithms (see [5] for an example) and will apply to all phase computations in this paper.

2.2 Time-Series from Dynamical Systems via Projections

It is common to define a sound synthesis algorithm through giving an equation or algorithm. Our purpose here is to clarify the relationships of different aspects of these equations in the light of general mappings from the circle to the circle. So we will look to give strong intuitions of the origins of the aspects of the algorithms in the light of the given topological domain. The formulation of circle maps allows to be precise about the relationship between stable topological space and the dynamics on it one the one hand, and the construction of the resulting time series on the other. Here we will discuss how we extract a time series y_n from an iterative dynamical system x_n. As an illustrative example, let us consider a constant step Ω around the circle [10, 14, 17][1]:

$$x_n = x_{n-1} + \Omega \mod 1 \tag{1}$$

[1] Throughout this paper the mod operation is applied to the whole expression.

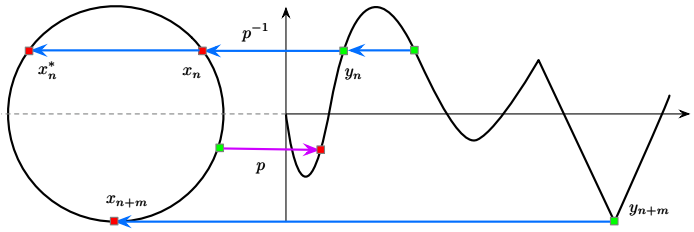


Fig. 1: Projection p and inverse projection p^{-1} between positions on a circle domain and a waveshaped time series. For most values the inverse projection is a two-valued function, hence an additional choice needs to be made to relate a given waveshape to a phaseshape.

Notice that there is a choice involved in how we treat this iteration as a time series. One is to simply interpret x_n as a sample and write $y_n = x_n$ leading to linear ramps. If we however consider sinusoidal choices such as:

$$y_n = sin(2\pi x_n) \qquad y_n = cos(2\pi x_n) \qquad y_n = sin(2\pi x_n + \psi)$$

Geometrically we can interpret these as three different choices of orthogonal projection of the motion on the circle onto a line domain (see Figure 1) and we recover sinusoidal oscillation. The rightmost case describes the general orthogonal projection from the circle under the choice of projection angle ψ. Using this intuition we will refer to mappings from a potentially higher dimensional dynamical system onto a discrete time line as *projection*, even if this mapping does not have a clear geometric meaning, and denote this function as $p(\cdot)$.

This is not the only choice of projection. A second projection that will be important in our discussion uses the phase function directly as output. It may appear trivial to explicitly point out the presence of a function that performs the identity operation $y_n = x_n$ but it serves the important recognition that this is a choice of how to interpret a dynamical system as time series. Some readouts of dynamical systems use linear transformations of the identity to rescale values and add offsets. For simplicity of discussion we will call all these cases *trivial projection* and we will use the symbolic notation $p(\cdot) = 1\cdot$.

Waveshaping as Modifying Projections: Waveshaping is the modification of a sine oscillator with a transfer function [1, 23]:

$$y_n = f(\sin(2\pi t_n)) \tag{2}$$

$f(\cdot)$ has some general restriction such as they it should be bounded between $(-1, 1)$ for inputs of $(-1, 1)$ in order to not exhibit gain. It is straightforward to interpret waveshaping as a composite projection $p(\cdot) = f(sin(2\pi\cdot))$. And an arbitrary choice of projection $p(\cdot) = f(\cdot)$ is a direct interpretation of the *phase-amplitude mapping* of a form of synthesis called *direct digital synthesis* in the context of digital hardware waveform generation [39]. A linear phase accumulator ϕ_n is directly fed into a wavetable lookup containing an arbitrary waveshape:

$$y_n = f(\phi_n) \qquad \phi_n : x_n = x_{n-1} + \Omega \mod 1 \tag{3}$$

Phaseshaping from Waveshaping: We can construct phaseshapes from waveshapes via the use of inverse projections. The goal is to find a phase-shaped function $f^*(\cdot)$ such that $y_n = f(p(x_n) = p(f^*(x_n))$ or in short $f \circ p = p \circ f^*$ (compare [17]). The specifics of this construction for an orthogonal projection is shown in Figure 1. Notice, that in most cases any back projection is a multivalued function. Except for normalized phases 0.25 and 0.75, there are two possible phases on the circle that can achieve any individual (forward) projection. To select a unique case, one can add a further criterion, such as the phase that is closest to a previous phase. The language of projection gives geometric meaning to this constructions, providing a template for generalizing these ideas in higher-dimensional cases, such as higher dimensional versions of phaseshaping [18].

2.3 Iterative Phase Functions

Next we seek to motivate *iterative phase functions*. When convenient we will abbreviate this to simply *phase functions* throughout this paper. These will be the main vehicle for connecting synthesis methods that have relations to oscillation. A simple definition of the iterative phase function is an iterative process that computes a discrete point x_n on a circle domain from a previous position on the circle x_{n-1} through some given mapping.

We have already encountered the simplest example of a phase function in equation (1). If we walk around a circle with constant phase steps Ω and project out through $sin(2\pi\cdot)$ we get the simple sinusoidal oscillator. Phase function (1) is sometimes called the *bare circle map* in the dynamical systems literature [14].

Synthesis algorithms are often given in terms of a time parametrization. To understand the reformulation from time-parametric to phase-iterative functions, let us again consider the sine oscillator with frequency ω over discrete time t_n:

$$y(t_n) = sin(2\pi\omega t_n) \tag{4}$$

We can convert this time-parametric version of the sine oscillator into a time-iterative version by writing a time step t_n as increment from a previous time step t_{n-1}:

$$t_n = t_{n-1} + \Delta t$$

To arrive at a phase iterative version, and hence the phase function for the sine oscillator of equation (1) we interpret our incremental changes as change in phase x_n with each iteration n. The frequency ω of equation (4) relates to phase increment Ω of equation (1) up to scale in time to phase dimensions. If a sample frequency ω_s and a fundamental frequency ω_0 are given, we can compute our phase increment as follows $\Omega = \omega_0/\omega_s$ [17].

From this individual mapping we construct a dynamical system by considering repeated iteration. The nth functional iteration $x_n = f(x_{n-1})$, along with some initial phase x_0, then is computed as follows (compare [36]):

$$x_n = f^n(x_0) = \underbrace{f \circ f \circ \cdots \circ f}_{n \text{ times}}$$

3 Oscillatory Synthesis Methods via Phase Functions

Next we will derive the phase functions for a range of oscillatory synthesis methods. In some cases, we will find that phase formulations are immediate, or have been provided in the prior literature. We will start with methods involving nonlinear feedback and follow with techniques absent nonlinear feedback such as modulation techniques. The cases of the sine oscillator has already been given in equation (1). Waveshaping and phase-amplitude mappings are not defined by the choice of phase function hence is treated in our discussion of projections in section 2.2.

3.1 Phase Functions with Nonlinear Feedback

Circle Maps: Circle maps have been proposed as candidates for providing chaotic oscillation for sound synthesis [10–12]. The most general form of circle maps refers to all mappings from the circle to itself [10]. Here we will restrict this to a perturbative form of the linear oscillator defined as follows:

$$x_n = x_{n-1} + \Omega + Hf(x_{n-1}) \mod 1 \tag{5}$$

$f(\cdot)$ refers to a nonlinear function. Throughout this paper a function is considered linear if it has the form $ax_{n-1} + b$ over the whole range of values $[0, 1)$ where $a, b \in \mathbb{R}$ are constants. All other functions we will call *nonlinear*. Hence, piecewise linear functions (linear only over a subset of $[0, 1)$), impulses, quadratic functions in x_{n-1}, trigonometric functions are all nonlinear functions. H is the strength of the nonlinearity[2] . If H is 0 then the map reduces to the phase function of the sine oscillator of equation (1). The choice of the nonlinear function $f(\cdot)$ provides a significant source of variation [11,12] not unlike general waveshaping. To distinguish this form of the circle map from those where a specific function $f(\cdot)$ has been chosen, we will refer to this form as *general circle maps*. It can be intuitively described as a one-parameter nonlinear perturbation of the sine oscillator.

Sine Circle Maps and Feedback Frequency Modulation: One of the most widely studied circle map uses a sine function $f(\cdot) = \sin(2\pi\cdot)$ as nonlinear perturbation and we will call this particular instance of the circle map the *sine circle map* [14]:

$$x_n = x_{n-1} + \Omega + H\sin(2\pi x_{n-1}) \mod 1 \tag{6}$$

It turns out that this phase function has appeared as a synthesis method, though the connection, as best as we know, has not been recognized so far. Consider the phase in the range of $[0, 2\pi)$ of the Feedback Frequency Modulation method [40, equation (1)], keeping the notation of the original:

$$y = x + \beta \sin y \tag{7}$$

[2] Much of the dynamical systems literature uses $H = -\frac{k}{2\pi}$ for the nonlinearity constant, as the sine circle map becomes non-invertible at $k \geq 1$ [10].

Clearly if we choose $\beta = H$, $x = \Omega$, and normalize to the range of $[0, 1)$ by dividing by 2π, this equation is identical to the sine circle map (6). For simplicity of the feedback in hardware the output projection of feedback FM was chosen to be $p(\cdot) = sin(2\pi\cdot)$. Hence we see that *feedback FM and sine circle maps are identical* when a sine projection is chosen, and they are closely related otherwise. Hence, we can immediately apply the literature on sine circle maps to understanding feedback FM.

Modulated Sine Circle Maps: A straight forward generalization of the sine circle map (aka feedback FM) is the introduction of a "modulation" frequency ω_m. This is not modulation in the sense of a fixed frequency, but a multiplicative change to a feedback frequency, which takes the same position as modulation frequency would in a construction without feedback:

$$x_n = x_{n-1} + \Omega + H \sin(2\pi\omega_m x_{n-1})) \mod 1 \tag{8}$$

It is a special case of the nonlinear perturbation by a Fourier series discussed in [11] where only one frequency is present.

Reciprocal Frequency Modulation as Additively Modulated Sine Circle Maps: Medine has proposed a method he calls reciprocal frequency modulation [26, Figure 4]. It differs from the modulated sine circle map by having a set modulation carrier that is perturbed additively by the feedback:

$$x_n = \Omega + H \sin(2\pi(\Omega_m + x_{n-1}) \mod 1 \tag{9}$$

In the language of circle maps this is an *additively modulated sine circle map.*

Functional Iteration and Nested Phaseshaping: Removing the term $x_{n-1} + \Omega$ and omit the the scaling factor H from general circle map of equation (5) reduces to the following form:

$$x_n = f(x_{n-1}) \mod 1 \qquad (10) \qquad x_n = \sin(2\pi\omega_m x_{n-1}) \mod 1 \qquad (11)$$

This approach has been introduced to sound synthesis by DiScipio under the name *functional iteration synthesis* (FIS) [36] and has been specifically studied using the sine function $f(\cdot) = \sin(2\pi\omega_m\cdot)$ [7, 36]. In his formulation it was not viewed as a phase function but was directly used as output. Hence we need the trivial projection $p(\cdot) = 1\cdot$ to retain our phase interpretation from the original proposal. A natural generalization is to use alternative projections. The projection $\sin(2\pi\cdot)$ is used in the context of nested phaseshaping [17], which is another generalization allowing the change of the phaseshaping functions between iterations. Hence, we get a more general understanding of iterated phaseshaping and functional iteration synthesis as nonperturbative nonlinear feedback iterations.

3.2 Phase Functions without Nonlinear Feedback

Phaseshaping: Phaseshaping are classical sound synthesis techniques [16, 17, 20, 22] and it can be generally written as follows:

$$y_n = \sin(2\pi f(\phi(t_n)) \qquad \phi(t_n) = \omega t_n$$

where $\phi(t_n)$ is a linear phase accumulator [39]. It is critical to note that t_n and y_n are independent in this formulation, hence there is no cascade or feedback in this construction. This can be formulated as phase function by introducing an independent iteration z_n as follows:

$$x_n = f(z_{n-1}) \qquad z_n = z_{n-1} + \Omega \mod 1 \tag{12}$$

Formulated as deviation from a sine oscillator, we get *perturbative phaseshaping* as follows:

$$x_n = x_{n-1} + \Omega + Hf(z_{n-1}) \qquad z_n = z_{n-1} + \Omega \mod 1 \tag{13}$$

Observe that if $z_0 = x_0$ and we set $H = 1$ then x_n becomes can be computed from a single function f that includes a additive term $+n\Omega$. Hence this first formulation and second formulation differs only in the content of the lookup table of the phaseshape. In equation 13 it contains the change in phase position relative to the previous phase position, whereas in equation (12) it contains the phase indexed by a linear phase accumulator, and these two formulations are closely related. The perturbative formulation allows a straightforward comparison to modulation techniques.

Frequency and Phase Modulation: Angle modulation techniques, of which frequency and phase modulation are closely related examples, are among the most successful oscillatory synthesis methods. A phase incremental formulation of frequency and phase modulation was given by Schottstaedt [35] that can be trivially unified into our notation for a sinusoidal modulation signal as follows:

$$x_n = x_{n-1} + \Omega + H\sin(z_{n-1}) \qquad z_n = z_{n-1} + \Omega_m \mod 1 \tag{14}$$

H is the modulation index. Ω_m is the phase increment associated with the modulation frequency. For general modulating functions $f(\cdot)$ we note that FM is a generalization of perturbative phaseshaping of equation (13), where the phase increment Ω_m of the perturbation is chosen independently.

4 Discussion and Generalization

Table 1 shows all oscillatory synthesis methods formulated with phase functions in the previous section. From it we can observe which changes transform one method into another. For simplicity we use $f(\cdot)$ whenever an unknown function can be used. It is important to keep in mind that these are not identical between different cases and methods and can have quite different interpretations.

Synthesis Method	Chaos	Projection	Phase Function $x_n, z_n = \ldots \mod 1$	Eqn
General Circle Map	✓		$x_{n-1} + \Omega + Hf(x_{n-1})$	(5)
Sine Circle Map	✓		$x_{n-1} + \Omega + H\sin(2\pi x_{n-1})$	(6)
Feedback FM	✓	$\sin(2\pi\cdot)$	$x_{n-1} + \Omega + H\sin(2\pi x_{n-1})$	(7)
Modulated Circle Map	✓		$x_{n-1} + \Omega + H\sin(2\pi\omega_m x_{n-1})$	(8)
Reciprocal FM	✓	$\sin(2\pi\cdot)$	$x_{n-1} + \Omega + H\sin(2\pi(\Omega_m + x_{n-1}))$	(9)
Functional Iteration	✓	$1\cdot$	$f(x_{n-1})$	(10)
Iterative Phaseshaping	✓	$\sin(2\pi\cdot)$	$f(x_{n-1})$	(10)
Iterated Sine Map	✓	$1\cdot$	$\sin(2\pi\omega_m x_{n-1})$	(11)
Phaseshaping		$\sin(2\pi\cdot)$	$f(z_{n-1}), z_{n-1} + \Omega$	(12)
Perturbative Phaseshaping		$\sin(2\pi\cdot)$	$x_{n-1} + \Omega + Hf(z_{n-1}), z_{n-1} + \Omega$	(13)
Frequency Modulation		$\sin(2\pi\cdot)$	$x_{n-1} + \Omega + H\sin(2\pi z_{n-1}), z_{n-1} + \Omega_m$	(14)
Sine Oscillator		$\sin(2\pi\cdot)$	$x_{n-1} + \Omega$	(1)
Waveshaping		$f(\sin(2\pi\cdot))$	$x_{n-1} + \Omega$	(2)
Phase-Amplitude Mapping		$f(\cdot)$	$x_{n-1} + \Omega$	(3)
General Modulated Circle Map	✓		$x_{n-1} + \Omega + Hf(x_{n-1}, z_{n-1}, \omega_m)$	(15)

Table 1: Phase functions and projections of oscillatory synthesis methods.

Furthermore we will choose the sine projection $p(\cdot) = \sin(2\pi\cdot)$ for sinusoidal oscillators, even if some other orthogonal projection is used (such as cos or some sine oscillation with phase offset). If the method is usually discussed without a specific projection in mind (as is typical for the dynamical systems literature), the entry is omitted from the table. Methods that are capable of exhibiting chaotic behavior are indicated in the table. The right-most column references the equation numbers from the discussion of these methods in this paper.

There are a few main differences between synthesis methods in Table 1. One difference is the absence or presence of the iterative constant phase increment $x_{n-1} + \Omega$ that corresponds to the simple sinusoidal oscillator from equation (1). Independent of the projection this term corresponds to a linear phase increase over time, also known as phase accumulator [39]. If another term is also present, then we can interpret the phase function as a *perturbation* to the linear phase case and under projection, as a perturbation to the sinusoidal oscillator. We see that a number of synthesis methods have been formulated perturbatively, while non-perturbative version has also been proposed. For example, Functional Iteration Synthesis (FIS) is a non-perturbative General Circle Map.

A second distinction is the use of the previous iteration x_{n-1} within a nonlinear function $f(\cdot)$ (which may be $\sin(\cdot)$) or alternatively some independent iterative constant phase increment $z_n = z_{n-1} + \Omega$. In the former case we have *feedback*, whereas in the former case the nonlinear function does not depend on the output. Notice that the presence of a past phase x_{x-1} is not sufficient here, as even the sine oscillator is computed from the past value of the phase. We see that for example the General Circle Map differs from Perturbative Phaseshaping only in the presence of feedback, as is the case for the difference of Phaseshaping and Functional Iteration.

The third distinction is the presence of a second frequency. Modulated Circle Map, Reciprocal FM, Iterated Sine Map, and Frequency Modulation have this property in our comparison. We note that this is a different property than an independent variable. If the independent variable z_n uses the same increment as the overall phase function it serves to protect the phase progression from the interference of a nonlinear function in the iteration. It does not introduce a second frequency.

Now we can write down the general form of an iterative equation that encompasses all methods discussed here by allowing each variation to be possible. It is important to note that this iteration is more powerful given that it allows arbitrary mixing between modulation and feedback aspects within an arbitrary function. We call this method *General Modulated Circle Map* to stay in line with other naming choices made:

$$x_n = x_{n-1} + \Omega + Hf(x_{n-1}, z_{n-1}, \omega_m) \mod 1 \tag{15}$$

Note, that we can convert this into a form absent the constant phase increment $x_{n-1} + \Omega$ by requiring that $\Omega = 0$ and that $f(x_{n-1}, z_{n-1}, \omega_m)$ contains the term $-\frac{1}{H}x_{n-1}$.

5 Conclusions

In this paper we have shown how many oscillatory synthesis methods can be understood as different instances of iterative phase functions that are broadly understood as mappings from the circle to itself, clarifying in particular when one should expect the possibility of chaotic behavior in the presence of nonlinearities. Furthermore it allows relating of these synthesis methods, understanding them as generalizations or special cases. We have shown that feedback FM and sine circle maps as near identical methods, we have given a phase perturbative version of phaseshaping, clarified the relationship of functional iteration and nested phaseshaping, and formulated a unifying generalization that include all discussed methods.

Two of the main components of our discussion, nonlinearity, and feedback are inextricably linked with Jean-Claude Risset's foundational research into sound synthesis methods as well as his later interest in chaos. More broadly we hope that this work gives an easier pathway to understanding the use of chaotic oscillation within the context of widely used oscillatory methods and provides a stronger connection between waveshaping, modulation type synthesis methods and chaotic dynamical systems. This paper can, however, be sensibly read as a proposal, that circle maps form a good candidate for inquiry into chaos in sound synthesis for its relationship to well-established synthesis methods and its intuitive formulation as perturbation of the sine oscillator.

There are numerous interesting avenues for future work. The author is planning a companion paper which provides quantitative comparisons of each algorithm with respect to parametric choices, given that the lack of space prevented it being included here. Finally, the development of concrete projection strategies for higher dimensional oscillators is another exciting topic of inquiry.

References

1. Arfib, D.: Digital synthesis of complex spectra by means of multiplication of non-linear distorted sine waves. Journal of the Audio Engineering Society 27(10), 757–779 (1979)
2. Berdahl, E., Sheffield, E., Pfalz, A., Marasco, A.T.: Widening the razor-thin edge of chaos into a musical highway: Connecting chaotic maps to digital waveguides. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 390–393 (2018)
3. Bidlack, R.: Chaotic systems as simple (but complex) compositional algorithms. Computer Music Journal 16(3), 33–47 (1992)
4. Chowning, J.M.: The synthesis of complex audio spectra by means of frequency modulation. Journal of the Audio Engineering Society 21(7), 526–534 (1973)
5. Collins, N.: Even more errant sound synthesis. In: Proceedings of the Sound and Music Computing Conference (SMC). vol. 6 (2012)
6. Di Scipio, A.: Composition by exploration of non-linear dynamic systems. In: Proceedings of the International Computer Music Conference. pp. 324–327 (1990)
7. Di Scipio, A.: Synthesis of Environmental Sound Textures by Iterated Nonlinear Functions. In: Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99) (1999)
8. Dobson, R., Fitch, J.: Experiments with chaotic oscillators. In: Proceedings of the International Computer Music Conference. pp. 45–48. Banff, Canada (1995)
9. Essl, G.: Mathematical Structure and Sound Synthesis. In: Proceedings of the International Conference on Sound and Music Computing. Salerno, Italy (2005)
10. Essl, G.: Circle maps as a simple oscillators for complex behavior: I. Basics. In: Proceedings of the International Computer Music Conference (ICMC). New Orleans (2006)
11. Essl, G.: Circle maps as a simple oscillators for complex behavior: II. Experiments. In: Proceedings of the International Conference on Digital Audio Effects (DAFx). Montreal (2006)
12. Essl, G.: Exploring the Sound of Chaotic Oscillators via Parameter Spaces (2019), under review.
13. Gogins, M.: Iterated functions systems music. Computer Music Journal 15(1), 40–48 (1991)
14. Hao, B.L., Zheng, W.M.: Applied symbolic dynamics and chaos. World scientific, 2nd edn. (2018)
15. Holopainen, R.: Self-organised sound with autonomous instruments: Aesthetics and experiments. Ph.D. thesis, University of Oslo (2012)
16. Ishibashi, M., et al.: Electronic musical instrument (Apr 21 1987), US Patent 4,658,691
17. Kleimola, J., Lazzarini, V., Timoney, J., Välimäki, V.: Phaseshaping oscillator algorithms for musical sound synthesis. In: Proceedings of the 7th Sound and Music Computing Conference (SMC) (2010)
18. Kleimola, J., Lazzarini, V., Timoney, J., Valimaki, V.: Vector phase shaping synthesis. In: Proc. of the Int. Conference on Digital Audio Effects (DAFx) (2011)
19. Kleimola, J., Lazzarini, V., Välimäki, V., Timoney, J.: Feedback amplitude modulation synthesis. EURASIP Journal on Advances in Signal Processing 2011(1), 434378 (2010)
20. Lazzarini, V., Timoney, J.: New Perspectives on Distortion Synthesis for Virtual Analog Oscillators. Computer Music Journal 34(1), 28–40 (2010)

21. Lazzarini, V., Timoney, J., Kleimola, J., Välimäki, V.: Five variations on a feedback theme. In: DAFx 09 proceedings of the 12th International Conference on Digital Audio Effects, Politecnico di Milano, Como Campus, Sept. 1-4, Como, Italy. pp. 1–7. Dept. of Electronic Engineering, Queen Mary Univ. of London, (2009)
22. Lazzarini, V., Timoney, J., Pekonen, J., Välimäki, V.: Adaptive phase distortion synthesis. In: DAFx 09 proceedings of the 12th International Conference on Digital Audio Effects, Politecnico di Milano, Como Campus, Sept. 1-4, Como, Italy. pp. 1–8. Dept. of Electronic Engineering, Queen Mary Univ. of London, (2009)
23. Le Brun, M.: Digital Waveshaping Synthesis. Journal of the Audio Engineering Society 27(4), 250–266 (1979)
24. Letowski, T.: Difference limen for nonlinear distortion in sine signals and musical sounds. Acta Acustica united with Acustica 34(2), 106–110 (1975)
25. Mackenzie, J., Sandler, M.: Modelling sound with chaos. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS). pp. 93–96 (1994)
26. Medine, D.: Dynamical systems for audio synthesis: Embracing nonlinearities and delay-free loops. Applied Sciences 6(5), 134 (2016)
27. Mikelson, H.: Mathematical modeling with csound: from waveguides to chaos. In: Boulanger, R. (ed.) The Csound book, pp. 369–384. MIT Press (2000)
28. Mudd, T.: Nonlinear dynamics in musical interactions. Ph.D. thesis, The Open University (2017)
29. Risset, J.C.: Catalog of computer synthesized sound. Murray Hill, Bell Telephone Laboratories (1969)
30. Risset, J.: Sound and music computing meets philosophy. In: Joint Proceedings of the 40th International Computer Music Conference, ICMC, and the 11th Sound and Music Computing Conference, SMC (2014)
31. Roads, C.: A tutorial on non-linear distortion or waveshaping synthesis. Computer Music Journal 3(2), 29–34 (1979)
32. Roads, C., Strawn, J.: The computer music tutorial. MIT press (1996)
33. Rodet, X., Vergez, C.: Nonlinear Dynamics in Physical Models: Simple Feedback-Loop Systems and Properties. Computer Music Journal 23(3), 18–34 (1999)
34. Schaefer, R.A.: Electronic musical tone production by nonlinear waveshaping. Journal of the Audio Engineering Society 18(4), 413–417 (1970)
35. Schottstaedt, B.: An Introduction To FM (2006), `http://ccrma.stanford.edu/software/snd/snd/fm`, retrieved on March 31, 2019.
36. Scipio, A.D.: Iterated nonlinear functions as a sound-generating engine. Leonardo 34(3), 249–254 (2001)
37. Slater, D.: Chaotic sound synthesis. Computer Music Journal 22(2), 12–19 (1998)
38. Suen, C.: Derivation of harmonic equations in nonlinear circuits. Journal of the Audio Engineering Society 18(6), 675–676 (1970)
39. Symons, P.: Digital waveform generation. Cambridge University Press (2013)
40. Tomisawa, N.: Tone production method for an electronic musical instrument (Feb 10 1981), US Patent 4,249,447
41. Truax, B.: Chaotic non-linear systems and digital synthesis: an exploratory study. In: Proc. of the International Computer Music Conference. pp. 100–103 (1990)
42. Valsamakis, N., Miranda, E.R.: Iterative sound synthesis by means of cross-coupled digital oscillators. Digital Creativity 16(2), 90–98 (2005)

Mathematics and music: loves and fights

Thierry PAUL

Centre de Mathématiques Laurent Schwartz, CNRS and Ecole polytechnique,
1128 Palaiseau Cedex, France
thierry.paul@polytechnique.edu
http://www.cmls.polytechnique.fr/perso/paul/

Abstract. We present different aspects of the special relationship that music has with mathematics, in particular the concepts of rigour and realism in both fields. These directions are illustrated by comments on the personal relationship of the author with Jean-Claude, together with examples taken from his own works, specially the "Duos pour un pianiste".

Keywords: music, mathematics, rigour, philosophy, Jean-Claude Risset.

Prelude: in memoriam

October 3rd 2016, I attended the Italian premiere of "Oscura". It was in Rome, with Maureen Chowning singing and John Chowning at the mixing board.

Unfortunately Jean-Claude couldn't come but I exchanged with him the following emails, alas the last ones.

Cher Jean-Claude,
je sors du concert au conservatorio à Rome,
où ton absence nous a surpris et ta presence
beaucoup manqué. Ton œuvre est très belle.
(...)
Très amicalement,
Thierry.
.

Cher Thierry,
Je te remercie vivement pour ton message de
Rome : j'avais escompté assister au concert
de Rome, mais j'ai dû rester à Marseille pour
des examens un peu énigmatiques.
Ton message amical m'a fait très plaisir.
Bien à toi,
Jean-Claude

1 Introduction

Jean-Claude Risset used to say: "[for example] *in music, we don't even have a Heisenberg principle, yet".*

This complain might seem strange, coming from a composer who used all along his life a lot of physics and mathematics in his action of composing. And on the other side, how not to imagine a bit awkward such a sentence in the mouth of a "pure" researcher other than Jean-Claude ?

This reference to the quantum uncertainty relations - which, by the way, is a (mathematical) theorem - is also very interesting if one thinks that quantum mechanics brought a deep change of paradigm in our way of seeing the world, which could be put in correspondence with the big changes we got from music - between the loss of tonality to nowadays music inseparable from computer sciences - in our way of hearing the sound [5].

But there is more in the relation between math and music, according to me. And much more than the link between the d'Arezzo notation and Descartes analytic geometry (much later), much more than group theory versus inversions, dilations, translations in the art of composing or improvising fugues (much earlier, again). Exactly like there is more than this analogy between wavelets and music scores, an analogy more pleasant to (some) mathematicians than to (any) musician.

For me, one of the deeper links between math an music, which also reveals fundamental differences, is located in the concept of rigour. Both mathematicians and musicians are rigorous in their action. No need to talk about the yoke that constitutes the obligation for the mathematician to prove things. Choosing a tonality, later a series, now a set of patches also consists in fixing some constraints which tie the composer in an a priori quite rigid frame. But this use, sometimes abuse, of rigour - and we claim that this is the same type or rigour which is truly used in both domains, a fact which characterizes them in between other sciences and art - is performed in very different places in music and in mathematics. As an a priori for the composer who feels free, in fine, to "cheat" with the constraints (in the same way with respect to tonality as to dodecaphonism). And, at the contrary, as a final achievement for the mathematician.

Rigour versus reality is another concept I would like to see math and music to share. In fact the famous (and for me a bit "has been") debate on mathematical Platonism - to put it in a nutshell, do mathematicians invent or discover theorems? - has some resonances in music: after all, any sound belongs (already(?)) to the nature, can one say naively. A comparison between (what one can call) realism in mathematics and music deserve, according to me, to be exhibited.

I will first discuss the position of rigour in both actions of doing mathematics and doing (composing) music: a priori for musicians, in fine for mathematicians. I believe Jean-Claude was a perfect illustration of this, and his works continue

to illustrate this way of handling this quite schizophrenic bridge between math and music.

Then I will briefly discuss the duality rigour/realism, and the tied concept of emergence in music and illustrate it by the "Duos pour un pianiste", this fantastic work by Jean-Claude where the unplayable rings real piano sounds.

2 Rigour in mathematics and in music: confluences and divergences

Music shares with mathematics a particular mandatory use of rigour. This is obviously, intrinsically true in mathematics. But it is also true for music, at least a certain kind of, as choosing tonality, series, a certain use of randomness, numerical patches, are constraint which constitute, according to me, a full tool box of rigour.

Composing, as proving, are actions very rigour consuming. But, at the contrary, the action of proving uses also a lot of other behaviours that escape completely from rigour. This is, this time, obvious for music, as everybody knows the disaster that created, in the past, a too rigid way of composing. But this is also true for mathematicians, who pass most of their time outside of rigour, being wrong.

Let us illustrate briefly these two common points for mathematics and music, namely the rigour and the loss of it, by the famous article "... wie die Zeit vergeht ..." by K. Stockhausen [11] and by looking at examples of writing (and I believe thinking) mathematics in an article by H. Poincaré published in the early twentieth century and a book by É. Goursat in the thirties (see [6] for more details.

In [11], Stockhausen makes the radical bet that one can compose by using in the same manner all the (very) different time scales (from the time of rythm (or even more, of the concert) to the one of timbre). This - in a certain sense quite natural - remark, very embedded in the rigorous serialism philosophy of music in the fifties, constitutes a starting point, a way of starting inside some very rigorous rules. Rules that the musician will be free to abandon during the process of composing.

At the contrary, in 1912, Poincaré [10] "defines" the famous Dirac δ function (a function equal to zero except at the origin where it takes an infinite value) in a significant but highly nonrigorous way. Strictly speaking the definition is empty. But it gives the whole flavour of what it should be, using a non defined notion of "infinitesimal small", a concept that Gourçat defines in a very floppy way[1] [4] twenty years later. In fact, one had to wait twenty years more to have

[1] "On dit qu'un nombre variable x a pour limite un nombre fixé a, ou tend vers a, lorsque la valeur absolue de la différence $x-a$ finit par devenir et "rester" plus petite que n'importe quel nombre positif donné à l'avance. Lorsque $a = 0$, le nombre x est dit "un infiniment petit".", Gourçat [4]

a rigorous definition of δ, much after its extensive effective use. On sees that, here, rigour enters the game after nonrigorous considerations, much after.

In fact the initial data are of the same type for the two fileds: an original material (a chord, a theme, a patch, an equation, a conjecture, an equality), but the way that have musicians and mathematicians to "honour their material" [2] are very different: limiting the imagination by strong constraints for the musician, fishing ideas outside any rigour for the mathematician. Eventually, the musicians will will free themselves from this rigorous straitjacket (sometimes quite quickly [3]) in the name of musicality and the mathematician will put some strict rigorous order in the different arguments necessitated by the constitution of the final proof.

Therefore, if mathematics and music share both the use of rigour and the loss of rigour, we arrive to the conclusion that there is a fundamental difference between them. In the action of doing mathematics and composing music, the musician start with rigour and eventually get rid of it by a typical artistic gesture. In the contrary of the mathematician, who starts outside any rules, fishing ideas, and end up with a perfectly rigorous situation produced by a typical scientific gesture[2].

In conclusion, musicians and mathematicians both need rigour, but at different times in their process of creation.

3 Rigour versus realism, and all that

What is real in mathematics, and in music? After having discussed the fundamental role of rigour in the processes of proving and composing, the aforementioned question concerns more the result of the production both in math and in music. Of course, the answer seems easier in music: what is real is the execution of the work during a concert.

There are no concerts in mathematics. Mathematical results are exchanged mostly through articles in specialized journals and discussions between experts. The diffusion of mathematics to non-mathematicians is something else, something apart, considered as non fundamental for the evolution of the field mathematics. What questions realism in mathematics is generally circumscribed to the notion of mathematical Platonism. Mathematical Platonism [8] states the problem of knowing if the contents of theorems are truly invented by mathematicians, as coming from a "nowhere", or, at the contrary, if mathematicians "just" discover their production in a "somewhere else" susceptible to contain everything.

There is a temptation [8] to consider that this "somewhere else" should be incarnated in music by the concert, a place of reality for music. But the situation seems to me less simple that it appears. Indeed: which concert? Which interpretation of the piece? And which "realization" in the the case of an open

[2] The reader interested in the concept of rigour in mathematics, philosophy and music, might consult the proceedings [1] of the conference RIGUEUR held in Paris, July 2 and 3 2019, to be published by Spartacus editions (Paris).

work? Placing realism in music too close to the acoustic event is, according to me, problematic.

I also strongly believe that the Platonistic debate in mathematics is a bit "has been" and that there is no "somewhere else": mathematics are just the result of their own construction. And I also think that realism in music doesn't seat in the place of concert: what is more real concerning the "Valses nobles et sentimentales"? The version for orchestra or the version for piano? What is real in "Duos pour un pianiste"?

In fact, putting realism in music inside the execution of a piece is by far too naive. It seems to me that realism emerge in music as a consequence of, among other things but necessary including it, the rigour present in the process of composing. reading a fugue without playing it is a real experience, after all. And conversely, considering that the mathematical realism calls only the rigorous part of mathematics is too simple, too reductive. Let us quote René Thom: "Ce qui limite le vrai, ce n'est pas le faux, c'est l'insignifiant"[3].

The conclusion of this short section could be expressed by saying that, both in music and in mathematics, the rigour has the role of providing a kind of "emergence of realism".

Acknowledgements. This work has been partially carried out thanks to the supports of the LIA AMU-CNRS-ECM- INdAM Laboratoire Ypatie des Sciences Mathématiques (LYSM).

References

1. Drouin, G., Paul, T., Rémy, B., Schmidt, M. (edts.): RIGUEUR. Spartacus, Paris, (2020). See https://indico.math.cnrs.fr/event/4602/ and http://www.cmls.polytechnique.fr/perso/paul/phenomath/.
2. Drouin, G.: Composer avec rigueur ou l'art d'honorer son matériau. In: [1].
3. Giavitto, J-L.: Formalisme, Exactitude, Rigueur. In: [1].
4. Goursat, É.: Cours d'analyse mathématique. Gauthier-Villars, Paris, (1933).
5. Paul, T.: Des sons et des Quanta. In C. Alunni, M. Andreatta, F. Nicolas (eds.) Mathématique/Musique/Philosophie. Collection "Musique/Sciences" IRCAM-Delatour, (2012).
6. Paul, T.: Rigueur-contraintes : mathématiques-musique. Gazette des Mathématiciens **139**, 71-77 (2014).
7. Paul, T.: platonisme - intrication - aléa (mathématique - physique - musique), à la mémoire de Jean-Claude Risset, magicien des sons impossibles. In: Proceedings of the conférence "Emergence en musique - dialogue des sciences". to appear.
8. Paul, T.: Mathematical entities without objects, on the realism in mathematics and a possible mathematization of the (non)Platonism - Does Platonism dissolve in mathematics?. European Review in press (2019).
9. Paul, T.: in memoriam. www.cmls.polytechnique.fr/perso/paul/inmemoriamtp.pdf

[3] What limits the true is not the false, it is the insignificant

10. Poincaré, H.: Sur la théorie des quanta. J. de Physique théorique et appliquée, 5ième série **2**, 5-34, (1912).
11. Stockhausen, K. : ...wie die Zeit vergeht... . Die Reihe, **3**, (1957). ...comment passe le temps... . Analyse musicale **6**, (1987).

Postlude: solo for a scientist and a musician

This conclusion will be an homage to Jean-Claude Risset, who all along his too short life, played continuously and successfully a permanent duo between him physicist and him musician [9].

"Duos pour un pianiste", which, together with this postlude, has a quite surrealistic but rigorous title, illustrate marvellously this duality. First of all, the work addresses the issue of the limits of virtuosity, a very musical one. These limits will be overcome thanks of the use of a computer, a evry mathematical object. Not by a computer creating electroacoustic sounds, but by a computer playing, through a precise and rigorous reaction to what the pianist just played, a piano Disklavier. And the only limit of virtuosity for the computer will be the one of the acoustic instrument.

Nobody knows really who, during the execution, is more influenced by the other: the pianist and the computer. Rather than a computer assisted piece of music, it is more of a mind assisted Nancarrow studies style piece.

Only the double-hatted mind of Jean-Claude could achieve such a miracle.

Exploring design cognition in voice-driven sound sketching and synthesis

Stefano Delle Monache[1] and Davide Rocchesso[2]

[1] Dept. of New technologies and Musical Languages, Conservatory of Music G. Verdi of Milano, Italy stefano.dellemonache@gmail.com
[2] Dept. of Mathematics and Computer Science, University of Palermo, Italy davide.rocchesso@unipa.it

Abstract. Conceptual design and communication of sonic ideas are critical, and still unresolved aspects of current sound design practices, especially when teamwork is involved. Design cognition studies in the visual domain represent a valuable resource to look at, to better comprehend the reasoning of designers when they approach a sound-based project. A design exercise involving a team of professional sound designers is analyzed, and discussed in the framework of the Function-Behavior-Structure ontology of design. The use of embodied sound representations of concepts fosters team-building and a more effective communication, in terms of shared mental models.

Keywords: Sound design · Cooperation · Design cognition.

1 Introduction

We witness an essential process of convergence of inquiries in sound design towards the broader field of design research. From different angles, sound studies, sonic interaction design (SID), computer science, auditory cognition studies, and sonification research are challenging the inherent cooperative and collaborative, yet ambiguous nature of listening and hearing, as method and means to contribute to better everyday environments for the living [2, 43]. More recently, sound design research has been unfolding its interest in the interaction- and information-centered use of sound in computational artefacts, towards the study of the process of designing sound. There are a variety of reasons to study sound designing. Researchers may want to have an understanding of the actual activities carried out by practitioners and their status [52]. Others may investigate the design process with the goal of improving the practice [12, 16]. Other loci of interest may inquiry designing sound with the aim of developing appropriate design tools and supporting technologies throughout the various stages of the process [7, 14]. Finally, other research approaches, whether bottom-up (e.g., case studies and design explorations [32, 44]) or top-down (e.g., reference frameworks and systems [50, 4]), may wish to achieve and provide a more general and abstracted explanation of thinking patterns in sound design tasks, the design process and methods that practitioners may look at.

Indeed, one main problem that sound practitioners strive to deal with is the communication and evaluation of a design, that is sound, in which the distance between the intermediary representations and the final product is perceived as very short. The intrinsic ambiguity of sound and listening do affect both the collaboration between peers and the contact points with stakeholder in general [10, p.35]: Communicating and elaborating concepts through sound can be hazardous, especially in the early interactions with clients; design solutions on sound are difficult to argue, especially when designers overindulge in the description of the sound-producing mechanisms rather than accounting for the global sensory experience; as a consequence, the evaluation of sound design proposals often takes the prosaic form of the individual preference of the client, whenever it is not based on psychoacoustic metrics for sound quality assessment [36, 40].

In the practice, sound creation and production rather unfold, within the overall design process, as an individual activity kept separate and asynchronous from the global product development, the effect of which undermines the participation and communication with stakeholders, both horizontally and vertically, and brings about a tendency to anticipate, early in the process, the creation of assortments of selected variations of highly refined sounds [46, 47].

Since the 2nd International Symposium "Les Journées du Design Sonore"[3] in 2004, a growing corpus of additive knowledge on sound design has been produced (see [24, 39, 49, 18, 37, 3, 43, 17], for a comprehensive overview). Such body of knowledge outlines a landscape of descriptive models of designing sound, whose central proposal is a closed loop of sound evaluation and design that advances through rapid prototyping and iterative improvement: Research through design workshops, design critique and explorations, and controlled experiments with sonic interactive artifacts and tools are primarily focused on the design activity in the conceptual and embodiment stages of the design process.

It turns out, however, that little is known about how sound designers think, generate and develop ideas. Given their multidisciplinary background [38], how do sound designers approach projects? For example, do they favor a search process in the problem space, like industrial designers do, or in the solution space, like engineers apparently prefer [28]?

Designing (sound) takes place in people's minds. Despite the market availability of countless types of computational tools for sound analysis and production, the very first creative act happens in the designer's mind. Understanding sound design thinking becomes crucial to create the next generation of design tools, computational or not, to aid the generation and communication of auditory concepts. In this respect, sound design research may look at design cognition studies in the visual domain, as reference framework of relevant topics, and rigorous and formal methods and ontologies to investigate sound design dynamics, individual and collaborative [11, 25].

[3] The symposium, organized by Frédérique Guyot (LAPS-design) and Patrick Susini (Ircam) in collaboration with the French Acoustical Society, took place at Centre Pompidou, Paris, France, in 2004 (https://www.centrepompidou.fr/cpv/resource/cazjxnn/rLLRyR).

In this area of study, protocol analysis is the established, empirical method, commonly used to inquiry well-defined design phenomena, such as novice-expert differences in problem structuring and organization of cognitive actions, the effect of the "structeredness" of ideation methods on cognition, the role of design representations and sketching, the conditions for design fixation and its effect on the novelty of ideas, and in general the cognitive processes involved in design moves [15]. Typically, audiovideo documentations of design sessions are transcribed and parsed in segments, that is the smallest units of analysis that can be time-based, reflect turn-taking (e.g., in team dynamics) or other rationale (e.g., decision-making), according the granularity and the objective of the study. Segments are coded according to meaningful schemes that may well-represent the particulars of the case in question. Finally, various kinds of qualitative and quantitative analyses can be carried out in order to derive an understanding of the design phenomena under scrutiny [23, 29].

In this work, we analyze the protocol of a design session involving a team of professional sound designers engaged in vocal sketching the sound of two car models (e.g., idle engine, driving, braking), with the aid of a computational tool for voice-driven sound synthesis [14]. We apply the Function-Behavior-Structure (FBS) ontology of design [29, chap. 13] as coding scheme, from which we derive information on the team dynamics and productivity, the role-taking, the designing style and process unfolding. The FBS ontology is a formal coding scheme which takes in account the cognitive processes emerging as transitions in the design space, and precisely in terms of transformations between classes of issues (i.e., the Function, the Behavior, expected and derived, and the Structure of the artifact under scrutiny) which are intrinsic to any design domain (see further, Section 3).

The paper is organized as follows: the next Section provides an overview of topics on cognition in conceptual design activity which can be relevant to investigate in sound design teamwork; Section 3 introduces the sound design session, and the coding procedure using the FBS ontology of design; we analyze and discuss the session in Section 4.

2 Conceptual design cognition at a glance

It has been argued that the next generation of CAD systems will be defined by four main characteristics, and namely cognition, collaboration, concepts, and creativity [20]. The majority of the research in design cognition concerns the disciplines of architectural design, engineering design, and product design, revolves around the two main paradigms of design as search in the problem space [21] or design as exploration and co-evolution [22, 48], and focuses on the processes of information gathering and structuring, the role of long-term memory, schema activation in working memory, semantic processing, mental synthesis and sketch-based reasoning (see [25, 26] for a systematic review of protocol studies on conceptual design cognition). These processes are essentially inspected from a visuo-

spatial perspective, if one excludes the role of the phonological loop for verbal design information [42, 8].

Such a structured knowledge and methodological approach are still missing in the realm of aural collaborative creativity and design. Certainly, auditory imagery and cognition represent a vast field of research which received an increasing attention in the past several years. The majority of the studies, originating from experimental psychology and neuroscience, are aimed at the understanding of the human ability to generate and manipulate mental auditory images, where music and language are the preferred foci of interest: For instance, it has been shown that auditory images contain both depictive and descriptive components, that is some relationships of the auditory stimuli are preserved (e.g., pitch and temporal properties), while others rather "sound like" (e.g., loudness), and that the reinterpretation (i.e., figure-ground segregation) of a given stream of sounds is more difficult in auditory imagery than in auditory perception, especially when subvocalization is blocked [27]. Functional MRI studies showed that conceptual acoustic processing, that is thinking about a sound even when implicitly presented through visual words, involves a partial reinstatement of the brain activity during the perceptual experience of the same acoustic features [31].

Empirical frameworks on the role of the active body as inherent mediator between perception and the cognitive processing of music (and sound) have been proposed, wherein the sonic experience emerges in interaction, as complex network of intentional states and internal models of observable patterns, that are acquired through knowledge and skills [35]. Within this framework, recent researches not only showed that vocal imitations of non-verbal sounds encode salient acoustic features into some other vocal features [33], while gestural metaphors are exploited to illustrate auditory sensations and causal representations of sonic concepts [34], but also explored their use as cognitive devices to enable and support sketch-based reasoning in conceptual sound design [13].

Embodied cognition, concepts, and creativity are at the center of frameworks for designing the next generation of sonic information and interactions [43, 45], where experimental applications of body-centered auditory display and sonification are finding their way in walking interactions in mixed reality, physical rehabilitation and motor learning, sensory alteration and emotional design [51]. On the other side, process-based studies on sound design are still embryonic [11], if one excludes the existing literature on creative thinking in music processes: The ill-defined problem of composing a piece of music is solved through iterative, non-linear stages of insight (i.e., musical inspiration), problem restructuring and proliferation [9]. Cognitive processes in music composition can reflect an analytic, horizontal approach to the sequential writing of the musical parts, or develop vertically, that is implying a strong conceptualization phase of the whole in advance, before the actual production: The "sound designer" style has been characterized by an in-depth, horizontal exploration and original use of tools, where iterations are especially concentrated in the re-execution and revision of sound segments in the recording phases [1].

Horizontal and vertical compositional strategies strongly resonate with the dual mechanism model of design problem-solving: lateral transformations of an idea are divergent and associative, are facilitated by ill-structured representations (i.e., conceptual sketches), and widen the problem space, while vertical transformations are convergent and inferential, are facilitated by well-structured representations (e.g., prescriptive sketches and blueprints), and deepen the problem space [21]. Sketch-based reasoning supports the re-organization and creation of new knowledge [5]. It can be argued that the main criticality of sound design is the unbalanced use of well-structured representations [12], due to the lack of proper tools that afford sketching in the established workflow, at least in the acceptation widespread in the visual domain [10, p.35].

Cooperation and collaboration in conceptual sound design are not common practices, where design in a natural setting, e.g. in a design firm, is typically carried out by teams with multidisciplinary background. The study of design teams may reveal several insights on design thinking. One main advantage of examining team design thinking is that think-aloud protocols are naturally enforced and concurrent with the task execution. It has been shown that early collaborations are the most effective and improve consistency in mental models, and yet that design teamwork do not necessarily brings about a higher productivity, compared to the individual activity: Apparently, the experienced individual designer is equipped with all the necessary expertise to act as a unitary system (i.e., a team, in which expertise are allocated by role-taking instead), where semantic coherence in team composition leads to better final design quality [15].

Design ontologies have been proposed to formalize and analize the design process. The Function-Behavior-Structure (FBS) coding framework was used to observe an industry team (a business consultant, three mechanical engineers, an electronic business consultant, an ergonomicist, and an industrial design student), involved in a brainstorming session: The distribution of word count and turns variation throughout the design episodes provided coarse quantitative observations on the quality of the team interaction, in terms of producing a shared mental model, where the analysis of transitions between FBS design issues produced fine-grained representations of the design process at the individual and team level [30].

The same coding scheme was applied to measure the designing styles of teams of industrial design students and mechanical engineering design students, that is observing whether the designers' focus on the problem space or on the solution space may be specific to design disciplines, and how it may affect team building and composition [28]: The problem-solution index was proposed as ratio measurement, computed over the total occurrences of design issues representing problem formulations (i.e., function and expected behavior) and design issues representing solutions to the formulated problem (i.e., structure, and behavior derived from the structure). The same measurement was applied to investigate how the structuredness of concept generation techniques affects the cognitive focus of teams towards the problem space or the solution space in the early stage of the design process [19].

Fig. 1. Team composition: project manager (P1), audio engineer (P2), sound designer (P3).

Taken together, this concise survey of relevant studies on team design cognition prompts a path of open questions in the domain of aural creativity and design. While the discipline of sonic interaction design has been proposing and accumulating a variety of methods, frameworks and techniques [4, 18, 13, 16, 6], to our knowledge no formal and structured inquiries have been carried out yet, in terms of their impact on sound design cognition. In the next Section we describe the sound design task, along with a brief, operational discussion of the FBS ontology, the coding procedure and conventions.

3 Cooperative sketching with voice-driven sound synthesis

The vocal sketching exercise took place during a sound design workshop organized in collaboration with the audio research team of a vehicle manufacturing company. The team, including two sound designers, two audio engineers, and one project manager, was split in two groups in order to derive two protocols that could be analyzed and compared.

The fictitious task was to re-design the EV (i.e., Electric Vehicle) engine sound of two car models, the Citröen 21DS rally and the Peugeot 205 GTI, according to three brand values, that is optimistic, smart, human. The overall design task duration was ninety minutes, split in forty-five minutes sessions per car model.

The teams were provided with a computational tool for voice-driven sound synthesis, to support their sketching activity: The tool affords the externalization of synthetic sound impressions by means of mixtures of sound models that can be set, played and shared as instances of vocal utterances. We refer to our previous work for the in-depth description and evaluation of the tool[4] [14]. The groups were also provided with silenced videos of the two car models, in order

[4] The tool in action: https://vimeo.com/271826511.

to video-prototype their design. The teams tackled the Citröen sound first, and the Peugeot sound design later. All the design tasks were video-recorded. In this paper, we report the protocol analysis of one team composed by one sound designer, one audio engineer and the project manager, as shown in Figure 1.

3.1 The Function-Behavior-Structure coding procedure

The FBS framework models any design and designing activity as a set of valid semantic transitions, that is cognitive processes, occurring between three classes of ontological variables (i.e., the purpose of the artifact, its imagined and emerging performance, the components of the artifact and their compositional relationships) that map onto six design codes that is issues, as shown in Figure 2: The final goal of designing is to transform a set of requirements (R) and functions (F) into a set of design descriptions of the artifact at hand (D).

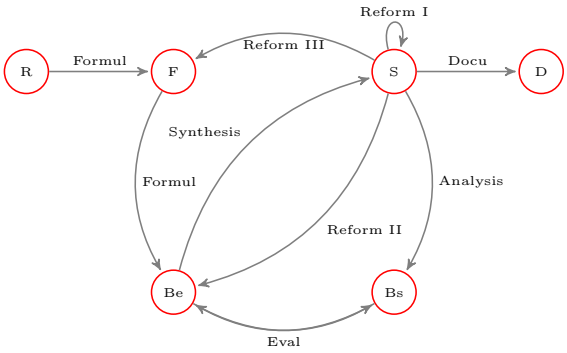


Fig. 2. The FBS framework. The codes represent the issues. Arcs are labeled according to the design processes.

In the practice, the transcript of the design session is parsed in design moves, that is segments of utterances, gestures, and any other kinds of representation, according to the classes of design issues. Design moves-issues are linked according to causal transformations. The designer formulates the function, based on the requirements (R $\rightarrow$ F), which are typically derived from the brief, or based on the Reformulation III driven by lateral thinking an existing structure (S $\rightarrow$ F). The expected behavior reflects the performance expected to fulfill the function (F $\rightarrow$ Be), and can be reformulated II by the structure (S $\rightarrow$ Be). The expected performance is synthesized in a structural configuration of elements and formal relationship (Be $\rightarrow$ S), that can be further inspected and revised (Reformulation I, (S $\rightarrow$ S). The analysis transition occurs once the structure is produced (S $\rightarrow$ Bs), and the actual performance based on the structure is assessed with respect to the expected behavior (Be $\leftrightarrow$ Bs). Eventually, this finite-state loop of design

processes among design issues leads to the documentation of external design descriptions (S → D).

In coding the transcript of the two design episodes (i.e., the Citröen 21DS rally and the Peugeot 205 GTI), we followed a set of conventions, extensively reported in [11], that we introduced to fully capture the sound designers' intentions, as they find themselves involved in discussing by means of verbalizations, vocalizations, iconic gestures accompanying the utterances, and synthesized sounds driven by either vocal control and manual control on the graphic user interface.

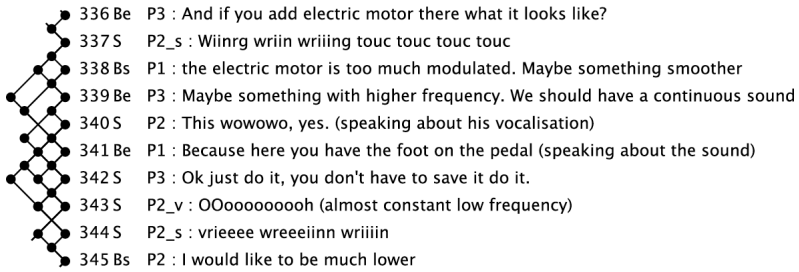


Fig. 3. Linkograph of an excerpt of 12 moves, characterized by an intense activity of synthesis (e.g., 336→337), reformulation of the structure (e.g., 340→344), analysis (e.g, 337→338) and evaluation (e.g, 338→339).

Once the coding procedure is completed, the emerging network of design moves can be organized and represented as a linkograph, where the relations among nodes, such as the density, the distance between links, the amount of links, their direction, distribution and patterns, are used to inspect the productivity and the reasoning of the designer(s) involved in the actual process [23]. Figure 3 reports the linkographic representation of a short excerpt of the sound design session, generated by feeding the software LiNKODER with the FBS-coded protocol [41]. In the next Section, we analyze and compare the two episodes, in order to derive a global picture of the team cooperation and dynamics.

4 Analysis and discussion

The Citröen sound episode (E1) resulted in a protocol of 444 moves, where 268 segments were retained for coding the Peugeot sound episode (E2). From the visual inspection of their linkographic representations, shown in Figure 4, we can notice that E1 is characterized by a longer link span (mean 10.5, STD 45.5), which denotes either longer incubation of ideas, where sketching serves as external memory function, and a typical team behavior where members may relate to her previous moves, regardless of the other moves intervened in the meanwhile. E2 linkograph is much more cohesive, with a shorter link span (mean = 5.3, STD = 10.3), which suggests a rather unitary behavior of team. This is strengthened by the E1 and E2 link indexes, a coarse indicator of the productivity

based on the ratio between the number of links and the number of moves, which are essentially comparable (E1, L.I. = 2.04; E2, L.I. = 1.97).

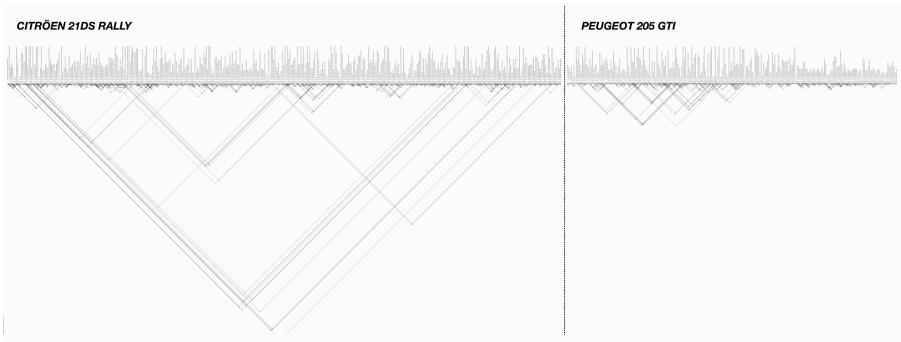


Fig. 4. Linkograhs of the two sound design episodes, based on the FBS coding.

The distributions of issues in episodes E1 and E2, reported respectively in tables 1a and 1b, do not show a statistical significance (Mann-Whitney, $U = 12$, $P > .05$). Similarly, the distribution of semantic processes is reported respectively in tables 2a and 2b (Mann-Whitney, $U = 17$, $P > .05$). The low percentage of D issues can be mainly attributed to the fact that only shared and agreed-upon documentations were coded. Taken together, the two episodes reveal a design activity especially centered on structural issues (S, Bs, $> 60\%$), where the cognitive efforts are mainly allocated to the reformulation of the structure and the evaluation of the resulting sound design. Two interpretation can be given: First, the intense work on structural issues can be ascribed to shortcomings of the vocal sketching tool, which hampered the production of synthetic representations coherent with the sketchers' intentions. This led the team to reconsider often the expected behavior, that is the imagined sound (S→Be, Reformulation II). Second, the team attitude is mainly addressed to the actual sound production, rather than its conceptualization (i.e., F→Be, Be→S). More in detail, tables 4a and 4b report the problem-solution index value for E1 and E2, which reflects the attitude of the team towards the design process, whether focused on the conceptualization of the design problem (P-S I. $> 1.$) or of the design solution (P-S I. $< 1.$). The shift to the diverse approaches may depend not only on the background, but also on the specific design task [28].

Tables 3a and 3b report the group activity in terms of use of verbalization, vocalizations and gestures, and externalized synthesized sounds. As expected, verbal-thinking is the main channel of communication. The different amount of use of other forms of communication by the team reflects the role and background of the members. The sound designer (P3) was apparently the most active in both episodes, where the audio engineer (P2) became less engaged in using vocalizations and gestures in E2. The project manager (P1) only interacted by talking.

CITRÖEN DS 21 RALLY

Total segments: **444**

ISSUE DISTRIBUTION %	
R	2 (0.5)
F	6 (1.4)
Be	122 (27.5)
Bs	114 (25.7)
S	195 (43.9)
D	5 (1.1)

Table 1a

SEMANTIC PROCESS DISTRIBUTION %	
Formulation	12 (1.5)
Synthesis	168 (20.8)
Analysis	137 (17.0)
Evaluation	164 (20.3)
Documentation	7 (0.9)
Reformulation I	196 (24.3)
Reformulation II	120 (14.9)
Reformulation III	4 (0.5)

Table 2a

TEAM ACTIVITY

	P1	P2	P3	TOTAL %
Verbalizations	67	77	141	285 (64,2)
Vocalizations	0	19	34	53 (11,9)
Gestures	0	28	18	46 (10,4)
Synth sounds	0	32	28	60 (13,5)

Table 3a: Individual and total occurrence

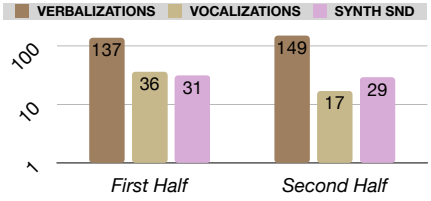


Figure 5a: Verbal and nonverbal expressionsions

PROBLEM-SOLUTION INDEX	0.42

Table 4a: Designing style focused on the problem (>1.) or the solution space (<1.)

INDIVIDUAL ACTIVITY

	- P1 - MANAGER	- P2 - ENGINEER	- P3 - DESIGNER	TEAM (TOTAL)
<CM4	1	8	18	27
CM4>	3	29	26	58
<CM4>	1	5	3	9

Table 5a: Backward, forward and bidirectional critical moves per participant

PEUGEOT 205 GTI

Total segments: **268**

ISSUE DISTRIBUTION %	
R	6 (2.2)
F	3 (1.1)
Be	87 (32.5)
Bs	77 (28.7)
S	92 (34.3)
D	3 (1.1)

Table 1b

SEMANTIC PROCESS DISTRIBUTION %	
Formulation	6 (1.3)
Synthesis	92 (19.9)
Analysis	74 (16.0)
Evaluation	125 (27.0)
Documentation	3 (0.6)
Reformulation I	94 (20.3)
Reformulation II	69 (14.9)
Reformulation III	0 (0.0)

Table 2b

	P1	P2	P3	TOTAL %
Verbalizations	58	54	87	199 (74.2)
Vocalizations	0	1	19	20 (7.5)
Gestures	0	4	22	26 (9.7)
Synth sounds	0	2	21	23 (8.6)

Table 3b: Individual and total occurrence

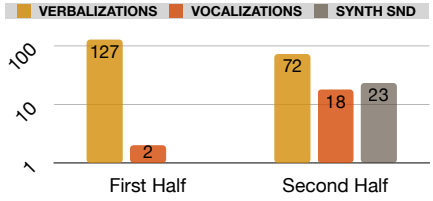


Figure 5b: Verbal and nonverbal expressionsions

PROBLEM-SOLUTION INDEX	0.56

Table 4b: Designing style focused on the problem (>1.) or the solution space (<1.)

	- P1 - MANAGER	- P2 - ENGINEER	- P3 - DESIGNER	TEAM (TOTAL)
<CM4	3	7	14	24
CM4>	8	8	15	31
<CM4>	1	2	8	11

Table 5b: Backward, forward and bidirectional critical moves per participant

However, the quality of the individual contribution in the two episodes is respectively reported in tables 5a and 5b: The critical moves (CMs) are moves with a high number of links, based on a significant threshold typically set around the $10 - 12\%$ of CMs of the total number of moves [23, p.73]. Links between moves are arranged in nodes of reasoning which may lead forward, thus denoting acts of synthesis (CM>), or backward, thus representing act of evaluation (<CM). Bidirectional moves (<CM>) are associated to rapid shift of divergent and con-

vergent reasoning. Critical moves represent turning points in the design process unfolding. Taken together, the team found a stronger integration and produced a more balanced process in E2 episode.

This is confirmed by the count of verbalizations, vocalization, and synthetic sound representations in the two halves of the protocols of the two episodes, shown in Figures 5a and 5b: Although the total percentage of verbalizations is even increasing, the two halves in E2 are very different in nature, where the drop of verbalizations and increase of nonverbal representations may stress the achievement of shared mental models, as a consequence of a better integrated conceptualization phase in the first half of the episode (see also the peculiar shape of the correponding linkograph in Figure 4). To conclude, the design exercise acted as an effective team-building tool, fostering a more effective communication between members with multidisciplinary background.

5 Conclusions

We showed how a cognition-based inquiry of sound design can reveal several aspects of team dynamics in conceptual design activities, which are considered critical and still unresolved in the current practices. The post-workshop comments further confirmed these findings. The participants reported a major frustration caused by several limitations of the computational sketching tool, that is first and foremost a lack of immediacy of use of the user interface and of technical integration in their established workflow (i.e., DAWs), where embodied interaction in the voice-driven sound synthesis prompted clear expectations on the creation and shaping of the sound sketches. Nonetheless, the participants remarked the high value of cooperation experienced in the embodied practice of sketch-thinking through voice-based representations of sonic ideas. They reported that for the first time, at least in their everyday workflow, they experienced to work collaboratively around a project. They especially remarked how through cooperation diverse approaches and ideas emerged. They found the overall workshop useful for team-building and reflecting on the role of creativity in their everyday sound design practice.

Designing is a process of construction of representations, from early, unstructured ideas of products, systems, services, etc. held in the mind towards the final artefacts. Protocol studies and ontologies of design provide established frameworks to tackle auditory cognition from a design mind-set perspective. Understanding how representations of sound designs are externalized for communication and collective transformation purposes becomes crucial to open sound design practices to truly participatory approaches, especially when users and stakeholders are involved not only as subjects, but especially as partners.

References

1. Barbot, B., Webster, P.R.: Creative Thinking in Music, pp. 255–273. Palgrave Macmillan UK, London (2018)

2. Barney, A., Voegelin, S.: Collaboration and consensus in listening. Leonardo Music Journal **28**, 82–87 (2018)
3. Barrass, S.: Sonic information design. Journal of Sonic Studies (2018), Special Issue on Sonic Information Design, https://www.researchcatalogue.net/view/558606/558686
4. Brazil, E.: A review of methods and frameworks for sonic interaction design: Exploring existing approaches. In: Ystad, S., Aramaki, M., Kronland-Martinet, R., Jensen, K. (eds.) Auditory Display - 6th International Symposium, CMMR/ICAD 2009, Copenhagen, Denmark, May 18-22, 2009, Revised Papers, Lecture Notes in Computer Science, vol. 5954, pp. 41–67. Springer (2010)
5. Brun, J., Le Masson, P., Weil, B.: Designing with sketches: the generative effects of knowledge preordering. Design Science **2**, e13 (2016)
6. Caramiaux, B., Altavilla, A., Pobiner, S.G., Tanaka, A.: Form follows sound: Designing interactions from sonic memories. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 3943–3952. CHI '15, ACM, New York, NY, USA (2015)
7. Carron, M., Rotureau, T., Dubois, F., Misdariis, N., Susini, P.: Speaking about sounds: a tool for communication on sound features. Journal of Design Research **15**(2), 85–109 (2017)
8. Cash, P., Maier, A.: Prototyping with your hands: the many roles of gesture in the communication of design concepts. Journal of Engineering Design **27**(1-3), 118–145 (2016)
9. Collins, D.: A synthesis process model of creative thinking in music composition. Psychology of music **33**(2), 193–216 (2005)
10. Delle Monache, S., Baldan, S., Boussard, P., Del Piccolo, A., Dendievel, C., Lemaitre, G., Lachambre, H., Houix, O., Rocchesso, D.: Interactive prototypes realized with the SkAT-VG tool. Deliverable, SkAT-VG project (2017)
11. Delle Monache, S., Rocchesso, D.: Cooperative sound design: A protocol analysis. In: Proceedings of the Audio Mostly 2016. pp. 154–161. AM '16, ACM, New York, NY, USA (2016)
12. Delle Monache, S., Rocchesso, D.: To embody or not to embody: A sound design dilemma. In: Fontana, F., Gulli, A. (eds.) Machine sounds, Sound machines - Proc. of the XXII CIM Colloquium on Music Informatics. pp. 93–100 (November 2018)
13. Delle Monache, S., Rocchesso, D.: Sketching sonic interactions. In: Filimowicz, M. (ed.) Foundations in Sound Design for Embedded Media, A Multidisciplinary Approach, pp. 79–101. Routledge, New York (2019)
14. Delle Monache, S., Rocchesso, D., Bevilacqua, F., Lemaitre, G., Baldan, S., Cera, A.: Embodied sound design. International Journal of Human-Computer Studies **118**, 47 – 59 (2018)
15. Dinar, M., Shah, J.J., Cagan, J., Leifer, L., Linsey, J., Smith, S.M., Hernandez, N.V.: Empirical studies of designer thinking: past, present, and future. Journal of Mechanical Design **137**(2), 021101–0211013 (2015)
16. Erkut, C., Serafin, S., Hoby, M., Sårde, J.: Product sound design: Form, function, and experience. In: Proceedings of the Audio Mostly 2015 on Interaction With Sound. pp. 10:1–10:6. AM '15, ACM, New York, NY, USA (2015)
17. Filimowicz, M.: Foundations in Sound Design for Embedded Media, A Multidisciplinary Approach. Routledge, London, UK (2019)
18. Franinović, K., Serafin, S.: Sonic interaction design. Mit Press (2013)
19. Gero, J.S., Jiang, H., Williams, C.B.: Design cognition differences when using unstructured, partially structured, and structured concept generation creativity tech-

niques. International Journal of Design Creativity and Innovation **1**(4), 196–214 (2013)
20. Goel, A.K., Vattam, S., Wiltgen, B., Helms, M.: Cognitive, collaborative, conceptual and creative-four characteristics of the next generation of knowledge-based cad systems: a study in biologically inspired design. Computer-Aided Design **44**(10), 879–900 (2012)
21. Goel, V.: Creative brains: designing in the real world. Frontiers in Human Neuroscience **8**, 241 (2014)
22. Goldschmidt, G.: The backtalk of self-generated sketches. Design Issues **19**(1), 72–88 (2003)
23. Goldschmidt, G.: Linkography: unfolding the design process. MIT Press (2014)
24. Grimshaw, M.: Game Sound Technology and Player Interaction: Concepts and Developments. IGI global (2010)
25. Hay, L., Duffy, A.H.B., McTeague, C., Pidgeon, L.M., Vuletic, T., Grealy, M.: A systematic review of protocol studies on conceptual design cognition: Design as search andexploration. Design Science **3**, e10 (2017)
26. Hay, L., Duffy, A.H.B., McTeague, C., Pidgeon, L.M., Vuletic, T., Grealy, M.: Towards a shared ontology: A generic classification of cognitive processes in conceptual design. Design Science **3**, e7 (2017)
27. Hubbard, T.L.: Auditory imagery: empirical findings. Psychological bulletin **136**(2), 302 (2010)
28. Jiang, H., Gero, J.S., Yen, C.C.: Exploring designing styles using a problem–solution division. In: Design Computing and Cognition'12, pp. 79–94. Springer (2014)
29. Kan, J.W., Gero, J.S.: Quantitative methods for studying design protocols. Springer
30. Kan, J.W., Gero, J.S., Tang, H.H.: Measuring cognitive design activity changes during an industry team brainstorming session. In: Design Computing and Cognition'10, pp. 621–640. Springer (2011)
31. Kiefer, M., Sim, E.J., Herrnberger, B., Grothe, J., Hoenig, K.: The sound of concepts: Four markers for a link between auditory and conceptual brain systems. Journal of Neuroscience **28**(47), 12224–12230 (2008)
32. Lemaitre, G., Houix, O., Visell, Y., Franinović, K., Misdariis, N., Susini, P.: Toward the design and evaluation of continuous sound in tangible interfaces: The spinotron. International Journal of Human-Computer Studies **67**(11), 976–993 (2009)
33. Lemaitre, G., Jabbari, A., Houix, O., Misdariis, N., Susini, P.: Vocal imitations of basic auditory features. The Journal of the Acoustical Society of America **139**(1), 290–300 (2016). https://doi.org/http://dx.doi.org/10.1121/1.4920282, http://scitation.aip.org/content/asa/journal/jasa/137/4/10.1121/1.4920282
34. Lemaitre, G., Scurto, H., Françoise, J., Bevilacqua, F., Houix, O., Susini, P.: Rising tones and rustling noises: Metaphors in gestural depictions of sounds. PloS one **12**(7), e0181786 (2017)
35. Leman, M., Maes, P.J., Nijs, L., Van Dyck, E.: What is embodied music cognition? In: Bader, R. (ed.) Springer Handbook of Systematic Musicology, pp. 747–760. Springer (2018)
36. Lyon, R.H.: Product sound quality-from perception to design. Sound and vibration **37**(3), 18–23 (2003)
37. Meelberg, V., Özcan, E.: Editorial: Designing our sonic lives. Journal of Sonic Studies (6) (2018), Special Issue on Sound Design, https://www.researchcatalogue.net/view/239747/239748/0/0

38. Özcan, E., van Egmond, R.: Product sound design: An inter-disciplinary approach? In: Undisciplined! Design Research Society Conference (2009), http://shura.shu.ac.uk/531/
39. Pauletto, S.: Perspectives on sound design. The New Soundtrack **4**(3), v–vi (2014)
40. Pedersen, T.H., Zacharov, N.: How many psycho-acoustic attributes are needed. Journal of the Acoustical Society of America **123**(5), 3163–3163 (2008)
41. Pourmohamadi, M., Gero, J.S.: LINKOgrapher: An analysis tool to study design protocols based on FBS coding scheme. In: Culley, S., Hicks, B., McAloone, T., Howard, T., Clarkson, J. (eds.) ICED 2011 Proc. of the 18th International Conference on Engineering Design, Impacting Society through Engineering Design. vol. 2: Design Theory and Research Methodology, pp. 294–303 (2011)
42. Purcell, A., Gero, J.S.: Drawings and the design process: A review of protocol studies in design and other disciplines and related research in cognitive psychology. Design studies **19**(4), 389–430 (1998)
43. Rocchesso, D., Delle Monache, S., Barrass, S.: Interaction by ear. International Journal of Human-Computer Studies (2019), accepted for publication
44. Rocchesso, D., Polotti, P., Delle Monache, S.: Designing continuous sonic interaction. International Journal of Design **3**(3), 13–25 (December 2009)
45. Roddy, S., Bridges, B.: Sound, Ecological Affordances and Embodied Mappings in Auditory Display, pp. 231–258. Springer International Publishing, Cham (2018)
46. Sander, H.: Listen! Improving the cooperation between game designers and audio designers. In: DiGRA '11 - Proceedings of the 2011 DiGRA International Conference: Think Design Play. DiGRA/Utrecht School of the Arts (January 2011)
47. Sanz Segura, R., Manchado Pérez, E.: Product sound design as a valuable tool in the product development process. Ergonomics in Design **26**(4), 20–24 (2018)
48. Schon, D.A.: The reflective practitioner: How professionals think in action. Basic Books (1984)
49. Serafin, S., Franinović, K., Hermann, T., Lemaitre, G., Rinott, M., Rocchesso, D.: Sonic interaction design. In: Hermann, T., Hunt, A., Neuhoff, J.G. (eds.) The Sonification Handbook, chap. 5, pp. 87–110. Logos Publishing House, Berlin, Germany (2011)
50. Susini, P., Houix, O., Misdariis, N.: Sound design: an applied, experimental framework to study the perception of everyday sounds. The New Soundtrack **4**(2), 103–121 (2014)
51. Tajadura-Jiménez, A., Väljamäe, A., Bevilacqua, F., Bianchi-Berthouze, N.: Body-centered auditory feedback. In: Norman, K.L., Kirakowski, J. (eds.) The Wiley Handbook of Human Computer Interaction, pp. 371–403. John Wiley & Sons (2017)
52. Zattra, L., Misdariis, N., Pecquet, F., Donin, N., Fierro, D.: Analysis of sound design practices [asdp]. research methodology. In: Fontana, F., Gulli, A. (eds.) Machine sounds, Sound machines - Proc. of the XXII CIM Colloquium on Music Informatics. pp. 168–175 (November 2018)

Morphing Musical Instrument Sounds with the Sound Morphing Toolbox

Marcelo Caetano[1] ⋆

INESC TEC - Sound and Music Computing Group, Porto, Portugal
mcaetano@inesctec.pt

Abstract. Sound morphing stands out among the sound transformation techniques in the literature due to its creative and research potential. There are several sound morphing proposals in the literature, yet few open-source implementations are freely available, making it difficult to reproduce the results, compare models, or simply use them in other applications such as music composition, sound design, and timbre research. This work describes how to morph musical instrument sounds with the sound morphing toolbox (SMT). The SMT is freely available and contains open-source implementations in MATLAB ® of a sound morphing algorithm based on sinusoidal modeling.

Keywords: sound morphing, musical instruments, sinusoidal model

1 Introduction

Sound morphing has found creative, technical, and research applications in the literature. In music composition, sound morphing allows the exploration of the sonic continuum [10, 14, 21]. Ideally, given the input sounds, sound morphing should allow automatically setting input parameters that control the morph to achieve the desired result [17]. Sound morphing is also used in audio processing, sound synthesis, and sound design. Tellman *et al.* [19] proposed a sound morphing technique based on sinusoidal modeling (SM) that is intended to improve the performance of a sample-based synthesizer by morphing between sounds of the same instrument to obtain intermediate pitches, dynamics, and other effects. Fitz *et al.* [6] use an SM called *Loris* to morph sounds. Sound morphing techniques have been used to investigate different aspects of timbre perception. Grey and Gordon [9] investigated the perceptual effect of exchanging the shape of the spectral energy distribution between pairs of musical instrument sounds. More recently, Carral [4] used spectral morphing to determine the just noticeable difference in timbre for trombone sounds. Siedenburg *et al.* [16] investigated the acoustic and categorical dissimilarity of musical timbre with morphing. However, these results are difficult to reuse, re-purpose, and build upon because seldom do we find freely available or open-source implementations of morphing algorithms.

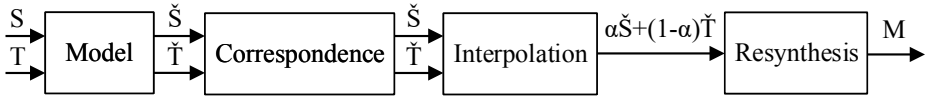
⋆ This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project: UID/EEA/50014/2019 and grant SFRH/BPD/115685/2016.

Currently, there are commercial morphing implementations available, such as Symbolic Sound's *Kyma* [1], SoundMorph's *Time Flux* [2], Melda Production's *MMorph* [3], and Zynaptic's *Morph* [4]. These commercial products typically have stable and bug-free implementations that can be controlled via a graphical user interface (GUI). However, besides the price, disadvantages such as little flexibility (i.e., control) and scarce technical information prevent their wider adoption in academic circles. A notable exception is *Kyma*, an implementation of the SM dubbed Loris [6]. However, composers and researchers alike need to be able to understand the algorithms employed and control several parameters of the transformation. There also exist closed-source implementations based on algorithms whose technical details can be found in publications. Ircam's *Diphone Studio* [5] uses the SM to morph between sounds. Trevor Wishart's *Sound Loom* [6] also allows morphing sounds. These are controlled via a GUI and the manuals typically contain little technical information because composers are the target user.

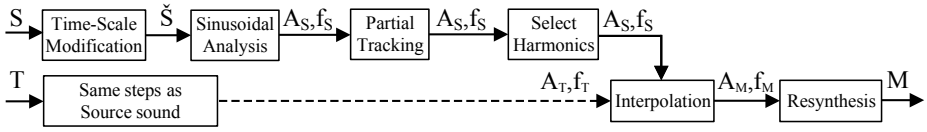
There are freely available open-source morphing implementations, such as Google Magenta's *NSynth* [7], Mike Brookes' *Voicebox* [8], and Hideki Kawahara's *STRAIGHT* [9] and *SparkNG* [10]. However, these find limited use in musical instrument sound morphing. NSynth uses a neural network synthesizer trained on a dataset with sounds from commercial sample libraries instead of recordings from acoustic musical instruments. Voicebox, STRAIGHT, and SparkNG were optimized for speech and their performance with musical instrument sounds remains untested. Dedicated sound models usually perform poorly on other sources.

This article describes the sound morphing toolbox (SMT), which contains MATLAB ® implementations of modeling and transformation algorithms used to morph musical instrument sounds. The SMT is open-source and freely available [11], making it highly flexible, controllable, and customizable by the user. The contribution of this work is the use of a practical example to show less technically inclined users (such as composers or researchers without the technical background) how to use the SMT. The next sections take the reader through the audio processing steps involved in morphing with the SMT, which are illustrated with figures and citations to the reference implementations. Section 2 presents an overview of the SMT and the source and target sounds used throughout the rest of the text. Section 3 shows the time-scaling algorithm, Section 4 describes the sinusoidal model used, Section 5 describes parameter interpolation, followed by re-synthesis in Section 6 and finally morphing in Section 7.

[1] `http://kyma.symbolicsound.com/`
[2] `https://www.soundmorph.com/product/24/timeflux`
[3] `https://www.meldaproduction.com/MMorph`
[4] `http://www.zynaptiq.com/morph/`
[5] `http://anasynth.ircam.fr/home/english/software/diphone-studio`
[6] `http://www.trevorwishart.co.uk/slfull.html`
[7] `https://magenta.tensorflow.org/nsynth-instrument`
[8] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`
[9] `https://github.com/HidekiKawahara/legacy_STRAIGHT`
[10] `https://github.com/HidekiKawahara/SparkNG`
[11] `https://github.com/marcelo-caetano/sound-morphing`



(a) General sound morphing procedure.



(b) The audio processing steps with the SMT.

Fig. 1: Overview of the morphing procedure in the SMT

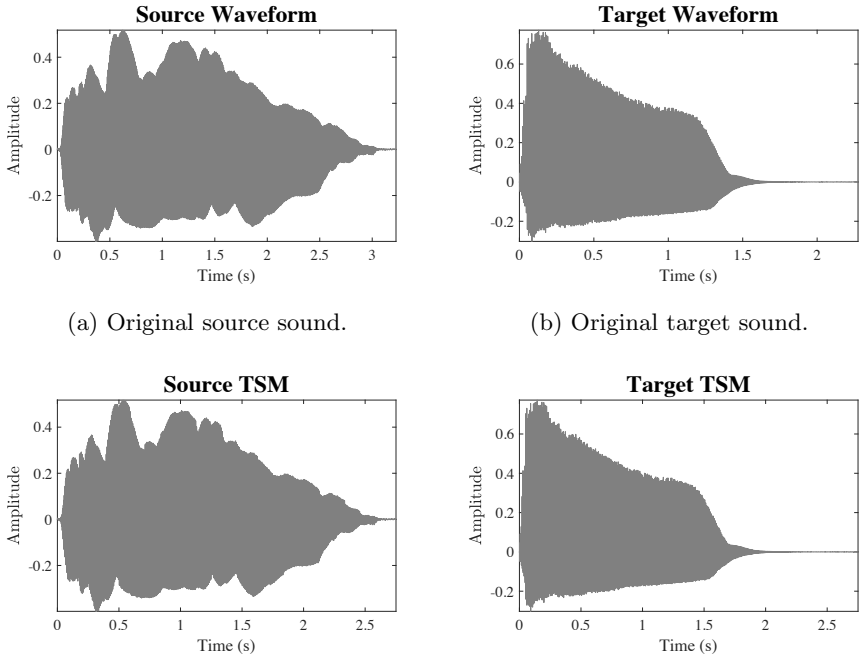
2 Overview

Figure 1 shows an overview of sound morphing with the SMT. Figure 1a shows the general morphing procedure and Fig. 1b shows the audio processing steps in the SMT. The SMT automatically morphs between a source sound S and a target sound T by setting the morphing parameter α that varies between 0 and 1. Only S is heard when $\alpha = 0$, whereas only T is heard when $\alpha = 1$. Intermediate values of α correspond to morphed sounds M with different combinations of S and T. For example, setting $\alpha = 0.5$ produces a morph that is halfway between S and T. Fig. 1a shows that, firstly, S and T are modeled to obtain a parametric representation $\check{S}$ and $\check{T}$. Next, correspondence between $\check{S}$ and $\check{T}$ is established, followed by interpolation and re-synthesis.

In Fig. 1b, we see a representation of the audio processing operations behind these steps in the SMT. First, both S and T are time-scaled to the same duration. Next, the SMT performs sinusoidal analysis of $\check{S}$ and $\check{T}$, producing the sets of parameters $\{A_S, f_S\}$ and $\{A_T, f_T\}$, namely, the amplitudes A and frequencies f of the sinusoids corresponding to $\check{S}$ and $\check{T}$. Correspondence in the SMT requires *partial tracking* and only the *harmonics* are interpolated because S and T are assumed to be nearly harmonic musical instrument sounds. The SMT establishes correspondence between harmonics of the same order and interpolates the amplitudes A and frequencies f using α to obtain $\{A_M, f_M\}$, which are used to synthesize the morphed sound M.

In what follows, the signal processing steps in the SMT corresponding to Fig. 1b are explained and illustrated with an example for $\alpha = 0.5$. Figure 2 shows the waveform of S and T used throughout the rest of the text. Fig. 2a shows S, a C#3 note played *forte* on an accordion. Fig. 2b shows T, a C3 note played *fortissimo* on a tuba. Listen to *Accordion_C#3_f.wav* and *Tuba_C3_ff.wav*. [12].

[12] Download the sound examples from `https://drive.google.com/file/d/1jHU-dGMTa_jXJ9LxPPnm-a3vV-yDbcO3/view?usp=sharing`



(a) Original source sound.

(b) Original target sound.

(c) Time-scaled source sound.

(d) Time-scaled target sound.

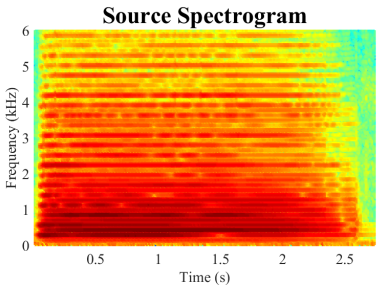
Fig. 2: Waveform of the original source S and target T and their time-scaled versions $\check{S}$ and $\check{T}$. S is "Accordion C#3 *forte*" and T is "Tuba C3 *fortissimo*".

3 Time-Scale Modification (TSM)

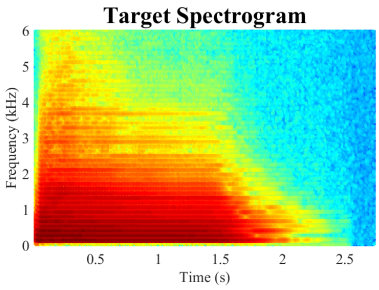
The first step is to use time-scale modification (TSM) [5] to establish temporal correspondence between S and T. In the SMT, the TSM algorithm implemented is *synchronized overlap-add with fixed synthesis* (SOLA-FS) [11]. SOLA-FS uses waveform similarity with an adaptable analysis step and a fixed synthesis step (see [11] for details). In the SMT, the morphing parameter α sets the final duration of the morphed sound (see [2] for details). Here, the duration of M will be halfway between that of S and T because $\alpha = 0.5$. Figures 2c and 2d show $\check{S}$ and $\check{T}$ respectively, which are S and T time-scaled. Note that the duration of both $\check{S}$ and $\check{T}$ is the same. Listen to *Accordion_C#3_f_tsm.wav* and *Tuba_C3_ff_tsm.wav* and compare with the original sounds. The next step is to use sinusoidal modeling (SM) to represent the oscillatory modes of $\check{S}$ and $\check{T}$.

4 Sinusoidal Modeling (SM)

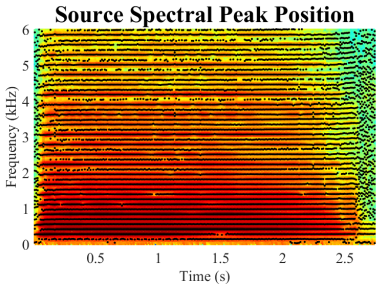
The SMT represents musical instrument sounds with the SM, which models a waveform as a sum of time-varying sinusoids parameterized by their amplitudes



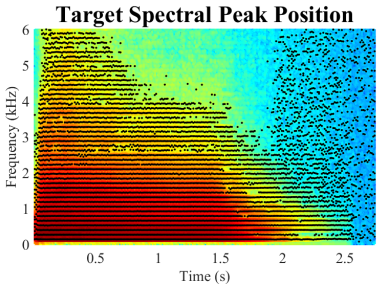
(a) Spectrogram.



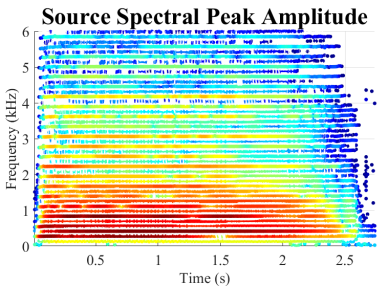
(b) Spectrogram.



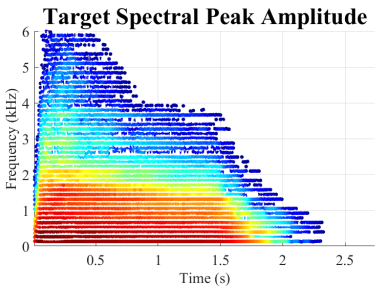
(c) Peak picking.



(d) Peak picking.



(e) Spectral peaks.



(f) Spectral peaks.

Fig. 3: Sinusoidal analysis with the SMT. The top row shows the spectrogram of S and T. The middle row shows the positions of the spectral peaks in frequency. The bottom row shows the frequencies and amplitudes of the spectral peaks.

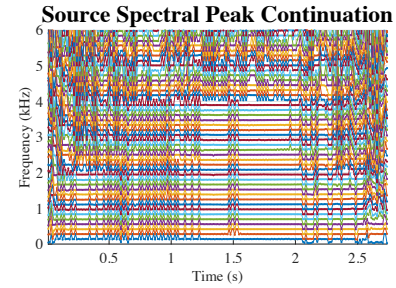
A, frequencies f, and phases θ [13, 15]. Sinusoidal analysis consists of the estimation of parameters, synthesis comprises techniques to retrieve a waveform from the analysis parameters, and transformations are performed as changes of the parameter values. The time-varying sinusoids, called partials, represent how the oscillatory modes of the musical instrument change with time, resulting in a flexible representation with perceptually meaningful parameters. The parameters completely describe each partial, which can be manipulated independently.

4.1 Sinusoidal Analysis

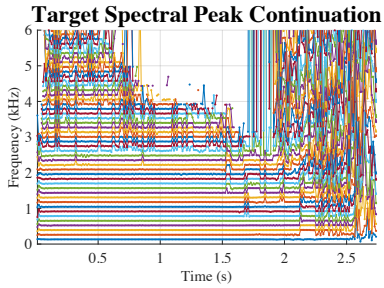
Sinusoidal analysis in the SMT uses spectral modeling [13, 15], which comprises *peak picking* and *parameter estimation*. The peaks of the magnitude short-time Fourier transform (STFT) are associated with underlying sinusoids whose parameters are estimated for each frame (and later connected across frames in the *partial tracking* step). The aim is to take the waveforms of $\check{S}$ and $\check{T}$ and, for each, output a set of time-varying sinusoids that correspond to $\check{S}$ (or $\check{T}$) when added together (hence the name *additive synthesis* [18]). The first step is to calculate the STFT. To ensure that harmonically related sinusoids can be properly resolved [18], the window length must obey $W \geq 3T_0$, where $T_0 = {}^{f_s}/_{f_0}$ is the period of the fundamental frequency f_0 (estimated with SWIPE [3] in the SMT) and f_s is the sampling frequency. In the example, C3 $\approx$ 131 Hz and C#3 $\approx$ 138 Hz and f_s = 44.1 kHz, so $W = \max\{945, 1010\}$ samples. A 1010-sample *Hann* window was used. The size of the DFT was $N = 1024$ (the next power of two after W), and the hop size $H = {}^{W}/_{2}$ (50% overlap).

Peak Picking The peak-picking algorithm in the SMT, described in detail in [13, 15], is illustrated in Fig. 3. The top row shows the spectrogram of $\check{S}$ and $\check{T}$, the middle row shows the position (i.e., the frequencies f) of the spectral peaks returned on top of the spectrogram, and the bottom row shows the spectral peaks with their corresponding amplitudes in dB (i.e., the final frequencies f and amplitudes A). Inside each frame, the SMT returns the $P_{\max}$ spectral peaks with the highest amplitude, so $P_{\max}$ sets the *maximum number of peaks* per frame. In Figs. 3c and 3d $P_{\max} = 80$. Additionally, the SMT allows to set the minimum relative amplitude level (in dB) of the final peaks with two different thresholds, namely the *local threshold* ρ and the *global threshold* ϱ. Inside each frame, only peaks that are at most ρ dB below the maximum are returned. Conversely, only peaks that are at most ϱ below the maximum level of the entire sound are kept. In Figs. 3e and 3f both ρ and ϱ were set to -66 dB. The effect of ρ and ϱ can be seen in the difference between Figs. 3d and 3f. In Fig. 3d we have $P_{\max} = 80$ peaks per frame, whereas in Fig. 3f peaks below ρ inside each frame and ϱ were removed (see the difference between 1.5 s and 2.0 s).

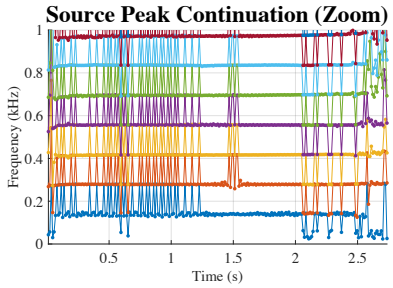
Parameter Estimation In the SMT, the values of the parameters of the SM (A, f, and θ) are estimated using either *nearest neighbor* estimation [13] or refined by *interpolation* [18, 15]. Both A and f can use quadratic interpolation



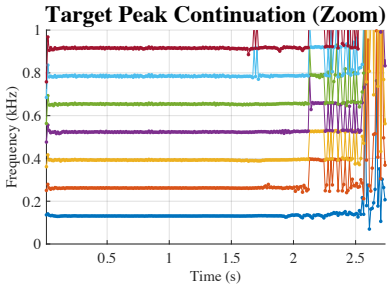
(a) Spectral peak continuation.



(b) Spectral peak continuation.



(c) Spectral peak continuation.

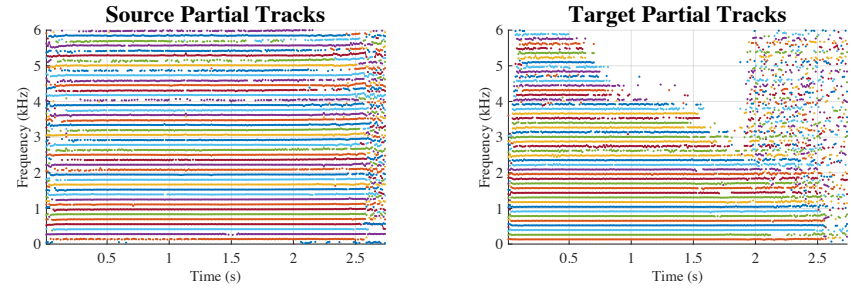


(d) Spectral peak continuation.

Fig. 4: Spectral peak continuation prior to partial tracking. The top row shows the spectral peaks connected by lines to illustrate the temporal discontinuity. The bottom part shows a zoomed-in part of the top row.

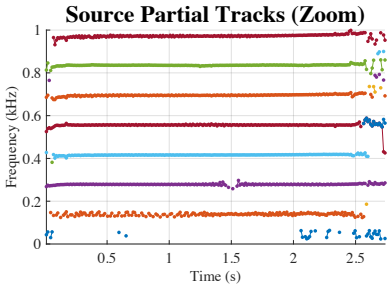
over a *linear* [15], a *logarithmic* [18], or a *power* [20] scale, whereas θ uses linear interpolation [18]. In Figs. 3e and 3f, the estimation of A and f was refined by quadratic interpolation over a logarithmic scale. The next section illustrates how to make the spectral peaks found into partial tracks. The phase values from $\check{S}$ and $\check{T}$ are discarded in the SMT. Section 6 provides details on how the morph M is re-synthesized via phase reconstruction by frequency interpolation [12].

Partial Tracking The spectral peaks returned from the peak-picking step (after further parameter estimation) do not result in a set of *continuous* partials because there is no mechanism to ensure temporal continuity. Figure 4 shows the final peaks returned from the parameter estimation step connected by lines. Inside each frame, spurious spectral peaks appear and later disappear (due mainly to interference by nearby peaks and sidelobe interaction), resulting in the discontinuous "spectral lines" in Figs. 4c and 4d. The SMT uses a *partial tracking* algorithm to convert the discontinuous spectral lines seen in Fig. 4 into the continuous partial tracks shown in Fig. 5. The partial tracking algorithm im-

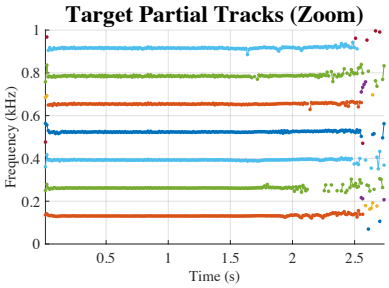


(a) Partial tracks.

(b) Partial tracks.



(c) Partial tracks.



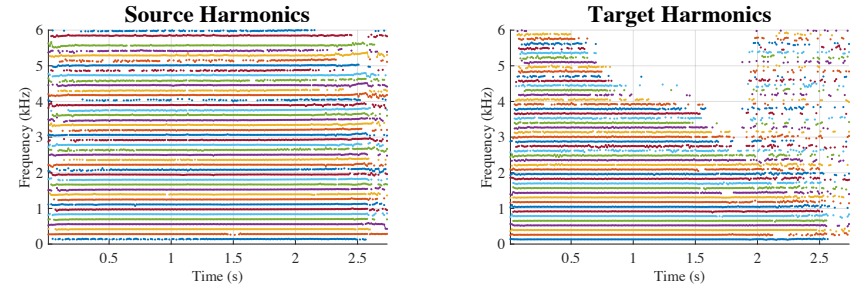
(d) Partial tracks.

Fig. 5: The result of partial tracking. The top row shows the spectral peaks from Fig.4 reorganized as partial tracks. The partial tracking algorithm ensures temporal continuity, as illustrated bu the zoomed-in regions on the bottom row.

plemented in the SMT is based on the peak continuation algorithm proposed by McAuley and Quatieri [13], so it simply collects peaks within a frequency threshold Δ_p into continuous tracks. In Fig. 5, $\Delta_p = f_0/4$, where f_0 is the fundamental frequency of $\check{S}$ or $\check{T}$. Note the difference between Figs. 4 and 5, especially the zoomed-in panels on the bottom. After partial tracking, the partials present continuous temporal trajectories, seen as fairly straight horizontal lines across. Listen to *Accordion_C#3_f_sin_part_ph.wav* and *Tuba_C3_ff_sin_part_ph.wav* to hear the result of re-synthesis using the original phase θ and all the partials.

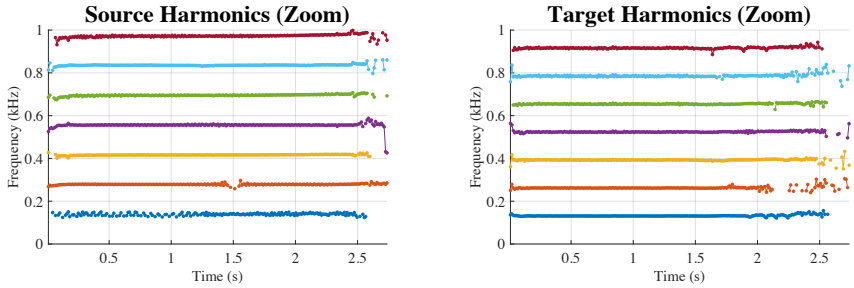
4.2 Harmonics

The next step after partial tracking is to select the harmonics of the fundamental frequency f_0. The *harmonic selection* step eliminates mainly the spurious frequency peaks while keeping the harmonically related partials, called *harmonics*. In the SMT, harmonics are the partials whose median frequencies over time are harmonically related to the fundamental f_0 within an interval Δ_h. This nearly harmonic relation can be expressed as $f_h = hf_0 \pm \Delta_h$, where f_h is the harmonic



(a) Harmonics.

(b) Harmonics.



(c) Harmonics.

(d) Harmonics.

Fig. 6: Only the partials whose frequencies are nearly harmonically related to the fundamental remain. Compare with Fig. 5. See text for details.

of order h. Figure. 6 shows the result of harmonic selection on the partial tracks from Fig. 5 with $\Delta_h = 10\,\text{Hz}$. Listen to *Accordion_C#3_f_sin_harm_ph.wav* and *Tuba_C3_ff_sin_harm_ph.wav* to hear the result of re-synthesis using the original phase and only the harmonics.

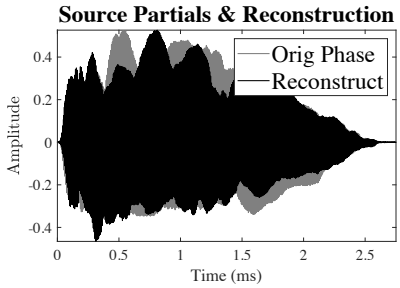
5 Interpolation

Prior to interpolation, the SMC establishes correspondence between harmonics using the harmonic number h. Then, frequencies are interpolated in cents as described in [2], whereas the amplitudes can be interpolated linearly or in decibels. The interval in cents c between two frequencies f_1 and f_2 is $c = 1200 \log_2 (f_1/f_2)$, so an intermediate frequency f_α is given by

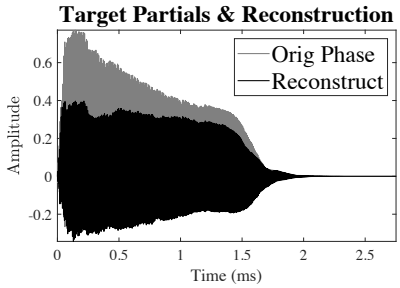
$$f_\alpha = \alpha f_1 2^{(1-\alpha)\log_2\left(\frac{f_2}{f_1}\right)} + (1-\alpha) f_2 2^{\alpha \log_2\left(\frac{f_1}{f_2}\right)}. \quad (1)$$

Similarly, for the amplitudes, the interval in dB between A_1 and A_2 is $\text{dB} = 20 \log_{10} (A_1/A_2)$ and an intermediate amplitude A_α is given by

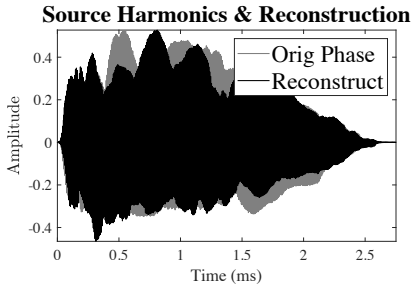
$$A_\alpha = \alpha A_1 10^{(1-\alpha)\log_{10}\left(\frac{A_2}{A_1}\right)} + (1-\alpha) A_2 10^{\alpha \log_{10}\left(\frac{A_1}{A_2}\right)}. \quad (2)$$



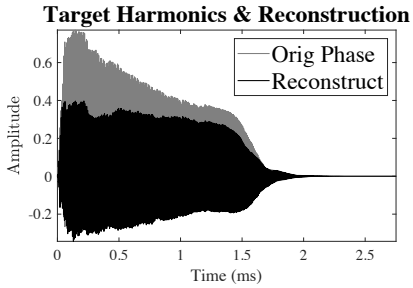
(a) All spectral peaks.



(b) All spectral peaks.



(c) Only harmonics.



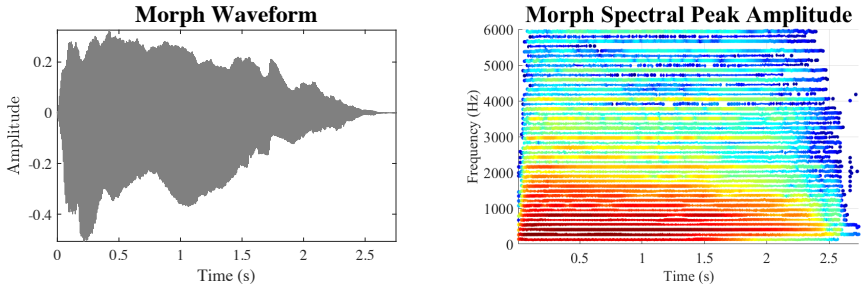
(d) Only harmonics.

Fig. 7: Comparison of different re-synthesis methods in the SMT.

Linear interpolation of amplitudes is achieved as $A_\alpha = \alpha A_1 + (1-\alpha)\, A_2$. In the example, linear interpolation of amplitudes was used. After all frequencies and amplitudes have been interpolated, the set of interpolated harmonics is re-synthesized to obtain the final morph.

6 Re-Synthesis

The SMT currently has three re-synthesis methods implemented, namely overlap-add (OLA) [8, 7], parameter interpolation (PI) [13], and phase reconstruction by frequency integration (PRFI) [12]. Both OLA and PI require the phases θ along with A and f to re-synthesize a waveform that is similar to the original both objectively and perceptually [18]. PRFI reconstructs the phase by integrating the frequency tracks across time [12], resulting in a waveform that is objectively different from the original but perceptually similar [18]. Figure 7 shows a comparison of the re-synthesized waveforms of S and T under four different conditions, namely all peaks (partials) or only the harmonics with the original phase θ (PI re-synthesis) and via PRFI. The top row shows both $\check{S}$ and $\check{T}$ re-synthesized using all spectral peaks. The bottom row shows re-synthesis



(a) Waveform of morphed sound.

(b) Spectrogram of morphed sound.

Fig. 8: Morphed musical instrument sound. The left-hand side shows the waveform and the right-hand side shows the spectral peaks with their corresponding amplitudes.

using only the harmonics. In all panels, the grey waveform uses the original phase θ (PI re-synthesis) and the black waveform uses PRFI. Listen to *Accordion_C#3_f_sin_harm_noph.wav* and *Tuba_C3_ff_sin_harm_noph.wav* to hear the result of re-synthesis using phase reconstruction and only the harmonics. Compare with *Accordion_C#3_f_sin_part_noph.wav* and *Tuba_C3_ff_sin_part_noph.wav*, which used phase reconstruction with all partials. Finally, compare these to their counterparts with the original phase θ.

7 Morphing

Finally, the morph is achieved by re-synthesizing the set of interpolated parameters M with PRFI. Figure 8a shows the waveform and the spectral peaks of the morph. Visual comparison between Fig. 8 and Figs. 2a and 2b does not seem to reveal the expected result. However, a comparison between Fig. 8a and Figs. 3e and 3f shows intermediate harmonics between $\check{S}$ and $\check{T}$. Listen to *morph.wav*.

8 Conclusions and Perspectives

This work has described how to use the Sound Morphing Toolbox (SMT) to morph musical instrument sounds. The audio processing steps were illustrated with figures and citations to the reference implementations. The SMT is open-source and freely available under a GNU3 license. Time-varying morphs [2] will be incorporated into a future version. Future development of the SMT will also add a GUI and an implementation of the hybrid source-filter model and the sophisticated sound morphing algorithm that uses it [1]. Finally, the SMT is currently an *alpha release* with possible bugs in the code due to limited testing. Adoption and use of the SMT by the community is encouraged to provide usability testing and bug corrections that might lead to a *beta release*.

References

1. Caetano, M., Kafentzis, G.P., Mouchtaris, A., Stylianou, Y.: Full-band quasi-harmonic analysis and synthesis of musical instrument sounds with adaptive sinusoids. Appl Sci **6**(5), 127 (2016)
2. Caetano, M., Rodet, X.: Musical instrument sound morphing guided by perceptually motivated features. IEEE Trans. Audio Speech Lang. Process. **21**(8), 1666–1675 (2013)
3. Camacho, A., Harris, J.G.: A sawtooth waveform inspired pitch estimator for speech and music. J Acoust Soc Am **124**(3), 1638–1652 (2008)
4. Carral, S.: Determining the just noticeable difference in timbre through spectral morphing: A trombone example. Acta Acust united Ac **97**, 466–476 (05 2011)
5. Driedger, J., Mller, M.: A review of time-scale modification of music signals. Appl Sci **6**(2) (2016)
6. Fitz, K., Haken, L., Lefvert, S., Champion, C., O'Donnell, M.: Cell-utes and flutter-tongued cats: Sound morphing using loris and the reassigned bandwidth-enhanced model. Comput. Music J **27**(3), 44–65 (Sep 2003)
7. George, E.B., Smith, M.J.T.: Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. IEEE Trans. Speech Audio Process. **5**(5), 389–406 (Sep 1997)
8. George, E.B., Smith, M.J.: Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. J. Audio Eng. Soc **40**(6), 497–516 (1992)
9. Grey, J.M., Gordon, J.W.: Perceptual effects of spectral modifications on musical timbres. J Acoust Soc Am **63**(5), 1493–1500 (1978)
10. Harvey, J.: "Mortuos Plango, Vivos Voco": A realization at IRCAM. Comput. Music J **5**(4), 22–24 (1981)
11. Hejna, D., Musicus, B.R.: The solafs time-scale modification algorithm. Tech. rep., BBN (1991)
12. McAulay, R., Quatieri, T.: Magnitude-only reconstruction using a sinusoidal speech model. In: Proc. ICASSP. vol. 9, pp. 441–444 (March 1984)
13. McAulay, R., Quatieri, T.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. **34**(4), 744–754 (Aug 1986)
14. McNabb, M.: "Dreamsong": The composition. Comput. Music J **5**(4), 36–53 (1981)
15. Serra, X., Smith, J.O.: Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition. Comput. Music J **14**, 12–24 (1990)
16. Siedenburg, K., Jones-Mollerup, K., McAdams, S.: Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. Front Psychol **6**, 1977 (2016)
17. Slaney, M., Covell, M., Lassiter, B.: Automatic audio morphing. In: Proc. ICASSP. vol. 2, pp. 1001–1004 (May 1996)
18. Smith, J., Serra, X.: PARSHL: An analysis/synthesis program for non- harmonic sounds based on a sinusoidal representation. In: Proc. ICMC. pp. 290–297 (1987)
19. Tellman, E., Haken, L., Holloway, B.: Timbre morphing of sounds with unequal numbers of features. J. Audio Eng. Soc **43**(9), 678–689 (1995)
20. Werner, K.J., Germain, F.G.: Sinusoidal parameter estimation using quadratic interpolation around power-scaled magnitude spectrum peaks. Appl Sci **6**(10), 306 (2016)
21. Wishart, T.: On Sonic Art. Routledge, New York, 1 edn. (1996)

Mapping Sound Properties and Oenological Characters by a Collaborative Sound Design Approach – Towards an Augmented Experience.

Misdariis Nicolas[1], Susini Patrick[1], Houix Olivier[1], Rivas Roque[1], Cerles Clement[1], Lebel Eric[2], Tetienne Alice[2], Duquesne Aliette[2]

[1] STMS Ircam-CNRS-SU
[2] Maison KRUG
nicolas.misdariis@ircam.fr

Abstract. The paper presents a specific sound design process implemented upon a collaboration with an important stakeholder of the wine (Champagne) industry. The goal of the project was to link sound properties with oenological dimensions in order to compose a sonic environment able to realise a multisensory experience during the wine tasting protocol. This creation has resulted from a large scale methodological approach based on the semantic transformation concept (from wine words to sound words) and deployed by means of a codesign method – after having shared respective skills of each field (sound and oenology). A precise description of the workflow is detailed in the paper, The outcomes of the work are presented, either in terms of realisation or conceptual knowledge acquisition. Then, future perspectives for the following of the work are sketched, especially regarding the notion of evaluation. The whole approach is finally put in the broad conceptual framework of 'sciences of sound design' that is developed and argued in the light of this study.

Keywords: sound design, codesign, taste, methodology, tools.

1 Introduction

The present project comes within the broad scope of crossmodal correspondences, i.e. the synesthetic associations that people tend to operate between different sensory modalities. In the literature, several studies aimed at describing or investigating the psychological mechanisms and the rationale of such associations [1]. The global aim of most of them is finally to understand how the percept in one modality can be interpreted or altered by sensory information given in another modality.

In the auditory domain, examples addressing that issue can be described, among others, with the following questions: what is the sound of a big/small or sharp/rounded objects? (e.g., the 'booba-kiki' effect studied by McCormick [2]). How does blue, red or yellow sound (e.g., the music/color correspondence, especially formalized by Kandinsky [3])? To what musical timbre a given flavor can be associated with (e.g., *bitter* paired with the French horn and *sweet* with the clarinet [4][5]). Or, quoting Charles Spence, can you taste the music? In other words, how music can influence the experience of taste, and change the personal emotional state?

In that domain, the "complex tasting experience that is drinking wine" [6] is especially focused. Not long ago, some works studied different interactions between what we drink and what we hear, among which the effect of different music styles on basic sensations (fresh, powerful, soft) during the taste of wine [7], or the perceptual and cognitive mechanisms underlying sensory modulations due to cross-modality [8].

In that scope, we recently conducted a long-term project (2017-18) with a famous French Champagne producer: Maison Krug. The goal of the project was to realize a mapping between sound properties and oenological characters in order to guide a sound design process and create an augmented multisensory tasting experience. This experience associates sound pieces with different types of wine coming from different regions. The challenge here was to understand the oenological concepts and transcribe them into sonic properties used, afterwards, for sound composition.

To do that, we worked in collaboration with members of the Krug winemaking team (Eric Lebel, the *Cellar Master* or *Chef de Caves*, and Alice Tetienne, *Winemaker*) and a music composer (Roque Rivas). Moreover, the project aimed at being implemented in a dedicated room – *La Salle des 400 vins* – where a specific multi-channel sound diffusion system has been designed and installed in order to render spatial properties of the sound production (see Fig. 1 and Sec. 4.2 for details).

Fig. 1. Krug tasting room (Reims) equipped with a 32-loudspeaker system (right, ©O.Warusfel), and the "400 wines wall" representing the 400 'vins clairs' held in the Krug wine library (left, ©P.Susini)

The main claim of this work concerns the concept of semantic transformation and one dedicated mean to achieve it, the collaborative design approach.

Semantic transformation is a concept that has been initially formalized in the visual domain by Karjalainen and Snelders [9]. It is a translation operation that addresses the issue of transcription of intentions. It relies on the association between words attached to given intentions (e.g. Brand identity) and words able to deliver design insights. Usually, semantic transformation are supported by mediations tools [10] like moodboards, card sets, etc. and are implemented within methodological frameworks.

Collaborative design – or *codesign* – can precisely be a relevant method to implement semantic transformation. It is a creative methodology based on a participatory approach that started to emerge in the late nineties [11][12]. It starts from the assumption that end-users are the experts of their own activity, so that they should be actively involved in the design process [13]. It is applied in several domains going from engineering to education, through design or arts [14].

In the sound design domain, semantic transformation and codesign have recently been studied and applied in a long-term research within a sound branding issue. In that frame, tools and methods were developed to convey sound identity and build corporate sounds [15][16]. That work aimed at making a link between the semantic definition of a Brand (*Brand-words*) and semantic descriptions of sound characteristics (*sound-words*), in order to provide sound design recommendations.

The present study is directly inspired from this process. It tends, this time, to implement a semantic transformation between oenological identities (*wine-words*) and *sound-words*, in order to give insight to sound design composition.

The article relates the workflow implemented to conduct this project. In a first stage, two expert groups (wine and sound) learnt from each other and passed on their respective knowledge and skills (Sec. 2). In a second stage, a codesign process is implemented and resulted in mapping strategies between *wine-words* and *sound-words* (Sec. 3). In a third stage, *sound-words* are transformed into composed sound pieces to illustrate oenological characters and transcend the tasting experience (Sec. 4). Then, after a conclusion, we open to perspectives, especially regarding evaluation, and finally reposition the whole project into a global conceptual framework.

2 Expertise Sharing

During the first stage, wine and sound experts learnt from each others. This stage was motivated by the participatory methodology implemented in the project. In fact, as the protagonists involved ought to work together within a collaborative framework, it appeared necessary to share a common expertise and language to speak about wines and sounds in order to elaborate efficient recommendations for the sound creation.

2.1 Speaking about wine

Vocabulary used to describe wine characteristics is quite huge. This is first due to the fact that wine tasting involves several sensory modalities corresponding to different sensory operations: we look at wine (sight), we smell it (smell) and we taste it (taste).

Each step brings specific information on wine. For instance, sight informs on color, intensity or viscosity (superficial aspects), smell informs again on intensity but also on flavors, and taste informs on mouth flavors (aroma, bouquet, etc.), balance or length in mouth. Each of these dimensions gets specific terminology and represents a semantic profile by itself. The visual analysis uses words like clear/blurry, brilliant/dull, fluid/thick, pale/intense, etc., together with all the shades of red (purple, burgundy, ruby, …) or white (colourless, yellow, golden, …) colors. The olfactory analysis uses words like closed/opened, poor/strong, etc. together with all the families of odors. The gustatory analysis uses words like soft/nervous, bitter, flexible/heavy, fleshy, velvet, short/long, etc. together with all the families of flavors.

This massively polymorphic character of wine comes mainly from the fact that it results from complex mechanisms (terroir, soil composition, sunshine, fermentation, conservation, etc.) occurring all through the production of the liquid that will become,

in fine, a wine or a Champagne. Precisely for Champagne – and especially in the Krug traditional process – this complexity is amplified by two elements: *i)* at early stage of production, a Champagne is a blend of several elementary wines – called '*vins clairs*' – that the Chef de Caves mixes together to build the Cuvées of the year; *ii)* vinification, and especially effervescence (formation of bubbles), takes at least seven years to be completed, a period during which the liquid inside the bottle goes on evolving according to its oenological nature.

On that basis, it was really ambitious and utopic to learn how to speak and taste about wine in the frame of the project – also considering that it takes a life of learning and practice to become a professional winemaker! Nevertheless, the Krug team made the task easy by, first, opening the doors of an internal tasting session and, second, delivering a simplified (but relevant) nomenclature of their oenological references.

The tasting session was a regular meeting of the Krug team (5 winemakers) dealing with the characterization of 15 yearly samples of 'vins clairs'. This kind of session is done twice a year: from October to December, after the grape harvest, and from February to March, before the Champagne creation. It is done for all the 'vins clairs' collected in that particular year, each of them corresponding to a specific grape variety ('*cepage*'), vineyard ('*cru*'), and even parcel ('*parcelle*'). All the tasting notes are registered in a repository document (the 'Krug *black book*') that also compiles notes from previous years. It is used during the blending process as a reminder of tasting notes (gustatory sensations) that have been previously produced.

The simplified nomenclature concerned the gross regions of growing included in the overall certified Champagne Region (East of France, around the city of Reims). In fact, after the first steps of the project, it was assumed with the Chef de Caves that this level of description was a good compromise between relevance (regions with indeed specific characters) and feasibility (number of regions compatible with sound design capability and sensitivity). Thus, 10 regions were defined (e.g. '*Montagne Reims Nord*', '*Cote des Blancs*', etc.) and specified with words in a 3-class typology: *i)* oenological cursors, i.e. six words that forms a standard grid for the oenological tasting at Krug (e.g., '*structure*' or '*expression*'); *ii)* additional terms, i.e. words that can be freely added by anyone in the winemaking team (e.g., '*roundness*' or '*liveliness*'); ii) additional marks, i.e. words that rather correspond to metaphoric associations or affective evocations. For that latter category, it is worth noticing that musical/instrumental metaphor were often used (e.g., violin, trumpet, marimba, etc.)

Outputs of this first section was then a table of 10 regions, each described by a group of words (20 in average) structured in 3 categories going from standard dimensions to free metaphors or associations, and a sound design team – including the composer R. Rivas – that was less novice about the semantic world and process of wine tasting.

2.2 Speaking about sounds

The second section consisted in the dual approach: learn novices in sonic issues (the Krug team) how to listen to sound and speak about them.

For that, we started from the research undertaken by Carron (2016) in the domain of sound branding and from which the present study is inherited [16]. In fact, based

on an analysis of several articles related to sound semantic description, Carron et al. (2017) proposed a list of common words related to sound features independent of the meaning, the process that produced the sound, or its location. Words are related to the sound itself, its acoustical characteristics and timbre features rather than illustrative analogies. Then, a lexicon – called afterwards *SpeaK* – that includes a list of 37 words was developed as an application displaying each word with a definition and sound examples in different categories (musical instruments, voices, environmental sounds, etc.). Within Carron's work, this operational tool was used as a training environment before a sound indexing task but also as a support for codesign sessions [17].

In the present work, the *SpeaK* tool was precisely used to introduce the wine experts (Krug team) to the world of sounds – and related words. Beforehand, the tool was improved in the light of the collaboration with the composer R. Rivas. A preliminary working session was organized in order to refine, and if needed extend, the lexicon.

This session actually gathered two composers (R. Rivas and Frederic Le Bel), three researchers (the 3 first authors of the article) and the sound designer who initially composed sound examples for the first version of *SpeaK* (Thomas Rotureau). During the session, the 22 semantic scales of the lexicon (15 bipolar scales + 7 single words) were methodically discussed with regard to the precision of the definition and the relevance of the sound examples. According to the latter point, the two composers were previously asked to prepare and bring for this session, alternative examples able to complement or improve the existing ones.

The global outcome of this session was a new release of *SpeaK* with 3 new elements: *i)* a simpler and user-friendly interface; *ii)* a 5-class generic structure of the sound examples (music, voice, environment, sound effects and basic synthesis); *iii)* if need be, new sound examples able to improve the quality of the illustration.

This current version of *SpeaK* was then used during a training session with the Krug team directly involved in the project: two members of the winemaking team – among whom the 'Chef de Caves' –, and the international marketing and communication director who also contributed to the collaborative process throughout the project.

The learning stage was inspired from Carron's experimental approach [16], and especially used the same training sound corpus. After some adjustments, the learning test included 4 exercises. An ownership period (45 min.) by a free browsing inside the lexicon. A first individual task (20 min.) that consisted in choosing one sound among five (5-AFC) with regard to a given semantic attribute – e.g. "choose the sound that is the most *fluctuant*". A second individual task (20 min.) that consisted in selecting one attribute among 10 words (a reduced list of the 37 words) with regard to a given sound. Answers from these two tasks were collectively discussed in order to share everyone's view. Finally, a third collective task (20 min.) consisted in the free description of a sound with 3 to 5 words from the lexicon. This last task allows the participants to agree on the perceptive qualities associated with a term in the lexicon.

Outputs of this second second learning section was, firstly, an adopted tool to support the semantic description of sounds, that will be used thereafter supporting the codesign session (Sec. 3), and secondly, a group of wine experts provided with expert language on sound which should refine perceptive capacities, enriches sensory exploration, and facilitates information selection, identification and comparison.

3 Collaborative Design

Then, after having investigated successively the wine and sound world (and words), we implemented a collaborative sound design (codesign) approach that basically consisted in going from the *wine-words* to the *sound-words* with a methodological process. This was done to give insights to the sound designer, so that he would be able to create and arrange the most relevant sound matters and forms.

3.1 Apparatus

The codesign environment was also transposed from Carron's work [16] and was formerly inspired by specific design approaches like *Kansei* [18]. Discussions were mediated by series of cardsets and the area of reflexion was materialized with a board, by analogy with a standard board game or role play. Supporting that, the *SpeaK* lexicon played the role of help to which anyone can refer during the session.

Three series of cardsets were built. They corresponds to the 10 wine regions, the 44 *wine-words* extracted from Krug terminology (6 oenological cursors + 38 additional terms) and the 37 *sound-words* of the lexicon, organized in 3 gross categories (General, Morphology, Timbre). Cardsets got a color chart to make their handling easier. Moreover, two things were added to facilitate the session proceedings: blank cards to possibly introduce new terms, and a trash bin to put aside non relevant words.

The board was configured with concentric circles considering that the bull's-eye (middle) was the region and its oenological description, and that the rim could be used to organise, and in case hierarchise, the corresponding sound properties (Fig. 2).

3.2 Protocol

After a warm-up stage (*SpeaK* browsing and exercise), the codesign session went as follows: one region is chosen, the Krug team fixed its oenological properties and placed the corresponding cards on the board. Then, semantic transformation is processed by discussing and placing cards around the target. At someone's request, examples or definition attached to a particular *sound-word* may be given by means of the lexicon. After a while, the instrumental metaphor is delivered to revitalize the debates and launch a new round of card handling. When all the participants agreed, the semantic portrait of the region is finalized and fixed (Fig. 2).

The processing of one region took approximately 30 minutes and systematically involved the same 7 persons. One of them, from the research team, played the role of mediator who opened the discussion and led or chased up the dialogue. The whole codesign stage (for 10 regions) occurred in 2 consecutive half-days.

At the very end, the semantic portraits were reconsidered to adjust descriptions in light of global coherence. Besides, few words were instinctively added to reach affective or emotional character of each region, such as 'serious', 'happy', 'warm', 'shy', etc. This additional contribution came out of the initial methodological frame, and was motivated by the composer who felt the need to collect more sensitive and complex dimensions than the basic sound properties given by the lexicon (Fig. 2).

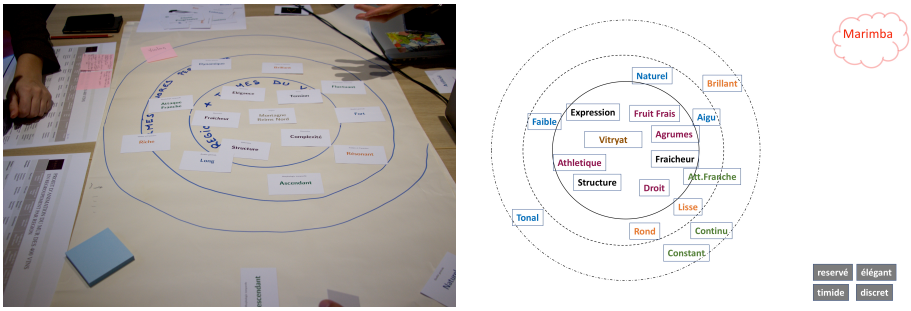


Fig. 2. Illustrative elements of the codesign sessions. The environment was formed of cardsets representing the different semantic spaces and board with concentric circles (left). At the end of each round, a semantic portrait of an oenological region is formed by associating wine-words – et the center – and sound-words – in concentric circles. In each portrait, the musical metaphor is recalled and some additional affective or emotional words are added (right). (©N.Misdariis)

3.3 Outcomes

The main outcome of the codesign stage was 10 semantic portraits respectively of the 10 oenological regions (Figure 2 gives an example on 'Vitryat'). From that, a synthesis was done, by trying to highlight global coherence and local differences among regions. For example, we tended to reveal global characteristics related to the Montagne Reims that were common to the 3 sub-areas (Reims Nord, Sud and Ouest), and local differences to discriminate between these same areas.

A reflexive look at this approach can also form another outcome of this stage. In fact, methodological elements can be usefully extracted from this experiment. They especially concern the role and the status of the mediator, the position of the composer – as the ultimate sound expert and potential session leader – or the use of extended sound examples, especially from musical / instrumental databases, able to feed the discussion and enlighten or consolidate raw ideas. These statements may certainly help us to improve the collaborative sound design approach which seems to be a rather specific practice within the general frame of codesign (see also Sec. 5.1).

4 Composition and Implementation

After the apprenticeship and codesign parts (Sec. 2 & 3), the third link of this project was mainly dedicated to composition and implementation of sonic transcriptions of concepts and words that have been handled, up to then. This part addresses two main issues: an artistic one dealing with the way to create sound sequences on the basis of words describing basic properties of sound or emotions; and a methodological one dealing with the way to transpose the work done in studio to the location where the sound design pieces were intended to be played – *a fortiori* if the sound diffusion system is technologically complex and massively multi-channels.

4.1 From description to composition

The composition started from all the semantic tags produced during the codesign session, plus the composer's personal notes. During the composition process (nearly 2 months), R. Rivas was also fed by some listening feedbacks (from the working team) that guided him for semantic interpretations and Krug's expectations on aesthetics.

The transcription work – from *sound-words* to sounds – is a challenge that contains a non-negligeable part of artistic intuition and, here, that sometimes forced the composer to read into root notions – *wine-words* or evocations – in a personal manner. Nevertheless, he tried to rationalize his approach by translating *sound-words* into the three fundamentals sonic dimensions: frequency, time and space.

Frequency and time were generally informed by the basic *SpeaK* vocabulary and gave rise to specific spectral contents or temporal envelopes. For instance, notions like 'low/medium/high', 'crescendo/decrescendo', 'brilliant/dull' or 'fluctuating' get nearly direct correspondences in the acoustic domain. Here, the main difficulty came from the need to combine several basic notions, make them physically compatible and musically relevant. Back to root notions, as evoked above, terms like 'complexity' (oenological cursor), or 'lemony' (additional term) can respectively be related to a dense multi-layer mix of several sound textures or high-pitched and rough elements – like the sound of the harmon trumpet mute that Miles Davis used to play.

On this point, the notion 'natural/artificial' used for some portraits also drove R. Rivas for the choice of specific sound synthesis paradigms. But for that, he also relied on the instrumental metaphor attached to each region by the Krug team (e.g., violin, marimba, horn, etc.). These essential references gave the color of each piece while staying subtile and integrated by means of sound transformations or complex mixing.

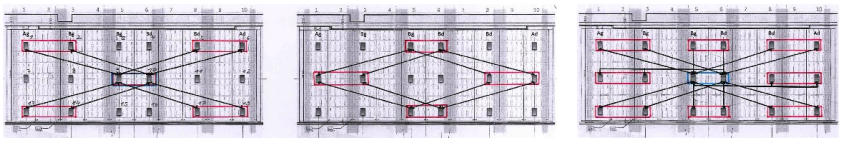


Fig. 3. Depiction of spatial sketches as stated by the composer R. Rivas in the composition phase. The background plane represents the '400 wines wall' and its 18-loudspeaker device and the drawings represent sound localisation or trajectories designed inside this grid.

Space was also fully used by R. Rivas to get another degree-of-freedom in the transcription process. In fact, thanks to the multi-channel device attached to the project, a real 3D-soundscape could have been designed in order to convey sound properties like 'close/far' or illustrate complex notions like 'aerial' or 'powerful'.

These spatial effects were able to be developed either in the frontal plane of the device (18-loudseaker wall) or in the surrounding space of the room by mainly dealing with opening, localisation, dynamic trajectories (Fig. 3), or in case, immersivity and envelopment.

4.2 From studio work *to location specific* mastering

The specificity of this project lies also in the sound diffusion system it is associated with. This is a 32-loudspeaker system, split in 3 parts: 3 lines of 6 speakers associated with the '400 wines wall', 6 loudspeakers placed in circle around the audience, 6 speakers integrated in the ceiling and 2-subwoofers (Fig. 1). This device – developped by Amadeus[1] upon design specifications made by the Ircam/EAC team –, offers a rather unique configuration for diffusing the sound pieces created in the project. But, added to specific acoustic conditions of the tasting room (glass or plaster walls, tiles floor), it also addresses sound engineering issues related to the mastering practice

This being, *in situ* setting sessions were set. They mainly consisted in: *i)* tuning the the audio mix quality, potentially altered by resonances or reflections due to room acoustics or electroacoustic response; *ii)* adjusting the 3D-soundscape with regard to the room dimensions and behavior. These operations were directly done in the audio sessions made in the studio and resulted in a multi-channel bounce for each of the 10 pieces, dedicated to be played back by a direct-to-disk device.

This part of the work points out a crucial practice in sound design: the *location specific* mastering. In fact, mastering is the final step of standard music production that aims at optimizing listening conditions in as many diffusion systems as possible. Sound design practices are commonly faced at specific mastering issues because of the diversity of sound devices usually used in this domain – from few (cheap) buzzers or loudspeakers placed in non conventional rooms (e.g. an automotive cockpit) to a lot of diffusion sources placed in large hall, such as a museum or a commercial hall[2]. One way to deal with this problem is to use audio simulation strategies based on 3D impulse response (IR) measurements that – by a deconvolution process – are able to render the effect of a given source diffused in the given room.

Presently, this virtual approach was not implemented for sake of time and project phasing. Instead, the composer and sound engineers (Clement Cerles and Colin Lardier), worked in studio either in a standard monitoring device or with a *pseudo*-'400 wines wall'. This system was built on purpose with similar unit sound devices and complied with the real volumic dimensions (wall's area, room's volume, etc.). In addition, as mentioned above, *location specific* mastering sessions were set and resulted in several modifications (EQs, internal balance of the mix, panoramic, tight filtering, etc.) that significantly improved the listening quality of the sound pieces.

5 Conclusions and Perspectives

To conclude, we conducted a long term project (nearly 2 years) that led us into the unexplored territory of wine industry and Champagne know-how. The main goal of the project was to design sound pieces informed by the knowledge of semantic correspondences between wine and sound worlds.

[1] http://amadeusaudio.fr/en/

[2] https://www.ircam.fr/article/detail/mastering-hors-du-studio-trois-experts-en-design-sonore-decryptent-des-nouvelles-pratiques-1/

The research and creation team was formed with researchers in auditory perception and sound design, a composer and a sound engineer who also played the role of a computer music designer [19]. The industrial collaboration was mainly interfaced with the Krug winemaking team and the marketing/communication department.

The project implemented the concept of semantic transformation and unfolded a participatory approach within a collaborative design (codesign) process. Moreover, it leaned on a methodological tool – previously developed and improved in the present frame: a sound lexicon (*SpeaK*) built as a dictionary collecting the major words used to describe sound properties, together with definitions and sound examples.

Within this frame, the project got four main stages in order to successively learn how to speak about wines, speak about sounds and collectively implement the semantic transformation from *wine-words* to *sound-words*. The fourth stage was dedicated to sound design (creation) and aimed at translating *sound-words* into sounds and musical composition that finally resulted in ten 1-min sound pieces diffused by a multi-channel sound device placed in a dedicated room: the Krug tasting room.

5.1 Perspectives

In the light of its originality and complexity, the Krug's project brings into front an emblematic approach. As a research process, it leads to open perspectives that should be further investigated in order to complement the project outputs. These perspectives mainly concern two components: codesign methodology and evaluation. As a sound design research, this project may also contribute to enhance our knowledge on the discipline and be part of a conceptual framework called sciences of sound design [20].

The codesign methodology applied to sound design seems to be quite encouraging and promising. As in its first implementation [16], the lexicon that supported the approach confirmed to be an efficient and relevant tool able to help communication and understanding on sounds. Nevertheless, the current codesign implementation showed some weakness that should be investigated and, may be, improved.

For instance, whereas the preliminary training exercises appeared to be unmissable, some uncertainties arose according the relevancy of the sound examples dedicated to these exercises and especially their ability to express just one basic property. Attention must be paid on the selection of these sounds and their polysemic content.

This precise issue is also addressed to the lexicon itself. In fact, on behalf of its rather 'encyclopedic' status, *SpeaK* must provide irrevocable and unequivocal specimens of sound examples illustrating sound attributes. This effort goes through the re-design of most of all examples from all categories. This work has yet started with the voice category by means of a recording campaign conducted by the composer R. Rivas; it will soon produce high quality and controlled vocal samples.

Indeed, concerning *SpeaK*, a more conceptual issue appeared during the Krug experiment: the fact that the list of pre-defined words were not sufficient enough to describe an oenological identity and that the composer needed high level descriptions (emotions, evocations, character) to be able to translate ideas into composition.

Finally, and more globally, we observed the fact that the experimental apparatus (board, cardsets, lexicon) prevented from describing dynamic changes of a semantic portrait as it often occurs in sound perception, but also wine tasting! This may force

us to imagine new paradigms that would also consider the temporal dimension of sounds which is not equally relevant in other domains (e.g. graphics).
On the other hand, whereas the PDS research group used to promote a 3-step sound design model (Analysis, Creation, Validation) [21], the evaluation stage is, right now, rather completely missing from the project proceedings – except few informal (and positive) feedbacks from the first tasting sessions at Krug. This point addresses an interesting and controversial issue that the project itself could help to investigate.

In fact, this asks the following fundamental questions: why should we evaluate and how can we proceed an evaluation? A rational answer to the first question could be: to verify the match between solutions and specifications, or to ensure the usefulness, usability, and desirability of the user experience produced by the solutions [22]. As for evaluation procedures, they should to be inspired, as usual, by the experimental psychology discipline with physiological, perceptual or cognitive measurements.

But, transposed to the current use case, the previous rationale appears to be more complex to argue and implement. In fact, in that case, the main specification to evaluate could be the perception of the wine characters into the sound composition. In other words, does one recover the basic oenological attributes of a region into a 1-min sound piece experience? Or more globally, is the semantic transformation finally valid, i.e. does it help the composer to create a relevant sound content and the listener to recognize the wine identity that intended to be illustrated? Or, alternatively, what does all of this bring to the tasting experience? All these questions address methodological issues in terms of experiment (what / how to measure?) but also in terms of contextualization (how to put the participants into controlled – and ethically acceptable – tasting conditions?). These issues form a work-in-progress reflexion that we hope to investigate and implement in a near future.

By listing all these outcomes and perspectives, we can observe that the Krug project brings considerable knowledge on the sound design discipline itself, its protagonists, its process and even its production. In that way, and even if it initially targeted a direct application – sound pieces composition for a tasting experience –, this project could finally be seen as a potential research project implementing a *research-through-design* approach that aims at producing knowledge instead of only solutions [23].

This precisely comes into the conceptual frame we recently tried to make emerge and promote, in accordance with the three research loci of Nigel Cross' formalization on design research: people, process and products [24]. In fact, transposed to the discipline of sound design, we look at laying the foundations of sciences of sound design that will investigate simultaneously the character of the sound designer, the sound design process, tools or methods and the status of the designed sound, i.e. what sound design produces *in fine* [20]. Then, to some degree, we can expect that the present project may have helped – and will help – to inform this approach and, by quoting Cross (2001) [25] give some elements to answer to the seminal question: is there a designerly way of knowing, thinking and acting in sound design?

Acknowledgments. The authors – namely the Ircam group – are sincerely thankful to Maison KRUG (Maggie Henriquez, and the last three authors of this paper) for their fruitful collaboration and flawless welcome. All authors would also like to thanks Cyril Beros, Jeremie Henrot (Ircam PROD), Olivier Warusfel (Ircam–STMS EAC), Gaetan Byk (Amadeus) and Emmanuelle Zoll for their contributions to the study.

References

1. Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971-995.
2. McCormick, K., Kim, J., List, S., & Nygaard, L. C. (2015, July). Sound to Meaning Mappings in the Bouba-Kiki Effect. In *CogSci* (Vol. 2015, pp. 1565-1570).
3. Kandinsky, W. (1971). *Du spirituel dans l'art*. Denoël/Gonthier.
4. Crisinel, A.-S., Spence, C. (2010). As bitter as a trombone: Synesthetic correspondences in nonsynesthetes between tastes/ flavors and musical notes. Attention, Perception, & Psychophysics, 72, 1994–2002.
5. Crisinel, A.-S., Spence, C. (2010b). A sweet sound? Exploring implicit associations between basic tastes and pitch. Perception, 39, 417–425.
6. Spence, C., & Wang, Q. J. (2015). Wine and music (II): can you taste the music? Modulating the experience of wine through music and sound. *Flavour*, *4*(1), 33.
7. North, A. C. (2012). The effect of background music on the taste of wine. *British Journal of Psychology*, *103*(3), 293-301.
8. Spence, C., & Wang, Q. J. (2015). Wine and music (I): on the crossmodal matching of wine and music. *Flavour*, *4*(1), 34.
9. Karjalainen, T. M., & Snelders, D. (2010). Designing visual recognition for the brand. *Journal of Product Innovation Management*, *27*(1), 6-22.
10. Boujut, J. F., & Blanco, E. (2003). Intermediary objects as a means to foster co-operation in engineering design. *Computer Supported Cooperative Work (CSCW)*, *12*(2), 205-219.
11. Kvan, T. (2000). Collaborative design: what is it?. *Automation in construction*, *9*(4).
12. Sundblad, Y. (2010, October). UTOPIA: Participatory Design from Scandinavia to the World. In *IFIP Conference on History of Nordic Computing*. Springer, Berlin, Heidelberg.
13. Darras, B. (2017). Design du codesign Le rôle de la communication dans le design participatif. *MEI Médiation et information, n 40: Design et communication*, 141.
14. Côté, N., Dubus, B., Fruleux, A., & Roche, C. (2014). Utilisation du Codesign dans la formation d'ingénieurs: exemple de projets en acoustique. In *Proceed. CFA 2014*.
15. Carron, M., Dubois, F., Misdariis, N., Talotte, C., & Susini, P. (2014, October). Designing sound identity: providing new communication tools for building brands corporate sound. In *Proceedings of the 9th audio mostly: a conference on interaction with sound* (p. 15). ACM.
16. Carron, M. (2016). *Méthodes et Outils pour Définir et Véhiculer une Identité Sonore* (Doctoral dissertation, Université Paris 6 (UPMC)).
17. Carron, M., Rotureau, T., Dubois, F., Misdariis, N., & Susini, P. (2017). Speaking about sounds: a tool for communication on sound features. Journal of Design Research , 15 (2)
18. Gentner, A., Bouchard, C., Badoil, A., & Favart, C. (2014, June). Kansei Cards: A Visual Tool Supporting the Investigation; Discussion; and Representation of the Kansei-Related Intentions of a Product to be Designed. In *KEER2014 Proceedings. Linköping; Sweden*.
19. Zattra, L., & Donin, N. (2016). A questionnaire-based investigation of the skills and roles of Computer Music Designers. *Musicae Scientiae*, *20*(3), 436-456.
20. Misdariis, N. (2018). Sciences du design sonore. Approche intégrée du design sonore au sein de la recherche en design. HDR thesis, Université de Technologie de Compiègne.
21. Susini, P., Houix, O., & Misdariis, N. (2014). Sound design: an applied, experimental framework to study the perception of everyday sounds. The New Soundtrack , 4 (2).
22. Robare, P. (2009). *Sound in Product Design* (Doctoral dissertation, Master thesis in Interaction Design. Pittsburgh, USA: Carnegie Mellon University School of Design).
23. Findeli, A. (2015). La recherche-projet en design et la question de la question de recherche: essai de clarification conceptuelle. *Sciences du design*, (1), 45-57.
24. Cross, N. (2006). Designerly ways of knowing . Springer London.
25. Cross, N. (2001). Designerly ways of knowing: Design discipline versus design science. Design issues , 17 (3), 49-55.

Kinetic Design

From Sound Spatialisation to Kinetic Music

Roland Cahen

Centre de Recherche en Design (CRD)
Ensci les Ateliers – Ecole Normale Supérieure Paris Saclay
roland.cahen@ensci.com

Abstract.
This paper explores the process of kinetic music design. The first part of this paper presents the concept of kinetic music. The second part presents the sound design and compositional process of this type of music. The third part presents some excerpts from the composition logbook of a piece called *Kinetic Design* to illustrate the process of kinetic design as work in progress.

This paper focuses on the question of sound spatialisation from a theoretical, as well as an empirical, point of view, through the experience and experiments of an electroacoustic music composer trying to make the imaginary concept of kinetic music real. It is a form of research by design, or research by doing. The kinetic design project examined here is the first time an experimental approach of research by design has been applied to kinetic music

Keywords: sound spatialisation, kinetic design, kinetic music, electroacoustic music, sound design, design process, composition

1 Introduction.

Kinetic music aims to produce sound choreography where the sound is diffused. Hence, it uses sound spatialisation in such a way that both the composer and the listener focus on kinetic aspects of sound, in opposition to using spatialisation only illustratively, or for rendering effects. In kinetic music, like theatre, dance or visual arts, each zone, position or direction can take on a musical value, a form of density that the sound space itself embodies. Kinetic music wishes to add a new form of expression and compositional methods to existing spatial music concepts and techniques. Sound spatialisation has already been the subject of abundant literature, the focus of this paper is to demonstrate the specificity of kinetic music. Of the literature on spatial sound, much has been written on the subject, such as generalities (principles, philosophy of space and music, history) numerous tools and techniques, analyses of musical intentions and abstractions about spatial figures, but very little has concentrated on the auditory experience, spatial sound aesthetics and none on the design process.

This article explores the hypothesis of kinetic music as a new process of composition and analyses its compositional process using my piece, *Kinetic Design*. The term design is used hereafter in relation to the creation process. The creative process explored here is 'research by design' or research by doing [6]. This form of research has a long legacy and is not novel in music. Indeed, Pierre Schaeffer developed his experimental approach by 'doing and listening' [1]. However, the kinetic design project examined in this paper is the first application of an experimental approach to kinetic music. By exploring the different motivations that drove the project, the successful and less successful experiments, and looking at extracts from my composition logbook, this paper tries to shed some light on some of the basic concepts and methods for kinetic sound design and composition. Kinetic Design was commissioned by INA-GRM, composed in octophony and performed on the acousmonium for the first time on January 20th 2019 at the MPAA Saint Germain (Paris).

2 The Characteristics of Kinetic Music

Kinetic Music aims to shape empty space with a choreography of sounds. However, a body of sounds would obviously be a kind of simulacrum[2], quite different from the human body or existing material objects. Composing and listening to kinetic music means focusing attention on spatial differences and similarities, as essential parts of kinetic musicality. Kinetic music adds new values to orchestration: spatial plasticity, incarnation and corporality, bringing an *orchestralisation*[3] of electroacoustic sounds. Kinetic music could be a breakthrough for electroacoustic/acousmatic music, anticipating new formal experiences and enabling new musical styles to emerge. However, to achieve these beautiful promises, the audibility of kinetic effects must be guaranteed.

2.1 Spatial Sound and Music Existing Work

There is a vast literature about sound spatialisation in the domain of music technology, electroacoustic music and perception. New rendering techniques procure nowadays a better sensation of sound incarnation thanks to new techniques such as WFS, High Order Ambisonics, dynamic pan and routing and room simulation. Simultaneously, spatialisation tools, editors such as the Ircam Spat, Panoramix, ICST Ambisonics, MaxMSP mc., and animation tools such Music Space, Acousmodules, Iosono Animix, GRM Tools (Spaces) and Iannix give most multichannel DAWs (digital audio workstations) and 3d real-time game editors (such as Unity3d) facilities for editing and automating sound source positions and motion. At the same time more international scientific, artistic and audio production centres are working on large multichannel sound devices to experiment with lines, arrays, matrix, domes and other sets of speakers. The existing

[1] Musical research approach by Schaeffer [11]

[2] Simulacrum refers to representation in ancient Greek literature and philosophy. This term was also used in XX century by Schaeffer [12]

[3] A neologism meaning making something orchestral which was not originally.

literature about sound spatialisation describes perception, concepts, descriptors and techniques.[4] Object based formats allow to give sound sources 3D(xyz) theoretical positions, that can be rendered to any standard multichannel soundfile format and in any real space, independently of the number and positions of the speakers. The SSMN Spatialization Symbolic Music Notation is an abstract and geometry toolkit for writing spatialised sound movements on instrumental scores. Spatial interactions with physical gestures are also explored at CIRMMT [5] and at Ircam[6].

2.2 Audibility as a Premise

Most spatial music rests on so-called trajectories, easily visible on sound editors' interfaces, but rarely audible in situ. However, spatial shapes, i.e. the compositional building block of kinetic music, should be clearly audible in order to be operational. Even if music can be heard when spatial sound shapes are not obvious, it is no longer kinetic music. What we could call *kineticality*, as a measurement of kinetic audibility, could be defined by how many audible elements of kinetic sound content or qualia[7] are lost in the spatial reduction process[8]. Unfortunately, musical audibility is more difficult to measure than individual sound-effect audibility. Spatial sound and music audibility can change a lot from one listening context to another and from one listener to another, for example because of the hotspot effect,[9] which some techniques, such as real source positioning[10] and linear panning[11] can reduce.[12]

Table 1. Easily Audible Kinetic Sounds/elements vs. Hardly Kinetic Ones

Audible	Inaudible (or less audible)
Recorded/synthesised impulses/attacks	Sine waves or tonic sounds with poor spectral complexity

[4] Blauert, Bregmann, Rumsey, Jot, Warusfel, Pachet/Delerue, Cadoz/Luciani, Brümmer, Duchenne, Pottier, Vandegorne, Schumacher, Schacher, Baalman, Orlarey etc.

[5] Schumacher with OMPrisma lib for open music

[6] Bevilaqua, Schnell, Lambert CoSiMa project.

[7] See the kinetic music qualia below.

[8] The reduction process consists of downmixing or converting multichannel sounds to stereo or mono.

[9] Only listeners placed in the hotspot position can hear the spatial effects properly, e.g. equidistant from the surrounding speakers.

[10] Real source positioning, as opposed to virtual positioning, techniques consist in placing each sound element on only one track / speaker. It has been used since the very beginning of spatialisation (multiple mono) and considered theoretically as a spatial compositional approach by various artists such as Pierre Boeswillwald in Octophonie Delta P (1991-94) with the concept of "octuor" or Benjamin Thigpen [13].

[11] Linear panning stands for panning between two next in line speakers, works also for a circle where the azimuth controls the position.

[12] The necessity to listen to spatialised sounds from a small point in the center of the device. The hotspot effect is more important when using virtual source positions than with real source positions.

Various noise dynamic modulations (filtering, granularity…)	Soft attacks are less precisely positioned
Tonal sound with rich complexity	Spatial complexity blurs kinetic sense
Reduced spatial complexity up to 2/3 voices	Complex movements

2.3 Kinetic Design Qualia

Here is a short list of kinetic qualities experienced in *Kinetic Design*. It is difficult to separate concepts, sensations or experiences from sound shapes.

- Immersion, position distribution, sound mobility.
- Punctual sounds and impulses: position (referent / changing), position blur, motion by re-iterations, motion by elementary transition, referent positioning, accentuation.
- Sustained sounds: motion by transitions (trajectory) between positions, spatial accentuation.
- Movements: pointed, traces drawn through space, rhythm, oscillation, swing, bouncing, hold and release, etc.
- Mass: extension / diversity, deployment/folding, spatial resolution.
- Plasticity, incarnation, sound corporeality.
- Explicit visual reference (visualisation) vs. implicit or metaphorical (materialisation, kinaesthetic shape, imagination).

In order to better understand where this typology comes from and how it has been established, section 3 will present some of the concepts of kinetic design qualia through the process of experimentation and composition, referring to the composition logbook of *Kinetic Design*.

2.4 Visual Imagery

Designing geometrical patterns of sounds, such as a sound trajectory, is not sufficient to produce kinetic sensation, but our imagination seems nevertheless to shape our representation of kinetic sounds. On the one hand, imagining these spatial sound shapes, our natural inclination is to project a visual grid on the auditory perception and to try to draw sound in space. However, by doing this, we might risk distracting the listener from the auditory experience. The act of hearing a sound should be sufficient in itself, and it is necessary to keep it at a distance from visual design, or tools for visualisation, so as to privilege audibility, i.e. auditory experience and sensation. On the other hand, "kinetic sound design" refers to visual imagery. Some visual metaphors are realistic and obvious, such as the big swing, or the little soldiers (Table 2) marching around the listener], others are more abstract, such as diversity variations or positional contrapuntal impulses (Table 2 Dots & Lines). This is why kinetic sound design belongs to the

discipline of design, since sound shapes emerge from the sound itself referring to our spatial representation of the world, which is mostly visual.
In this article the term design is also used in relation to the creation process which this article presents through a "research by design' approach.

3 Composition Device, Design Process and Workflow

A description of a kinetic design and compositional process follows. This enquiry encompasses theoretical and then applied research. Since section 4 contains a presentation of my experience and methods of kinetic design composition, as a concrete illustration of this process, an outline of the composition follows in order to help the reader situate the different sections of *Kinetic Design* discussed below. The workflow depicted below is only one among others, but it focuses on i) kinetic audibility and ii) spatial composition and spatial authoring possibilities at every stage of the compositional process from its conception to performance.

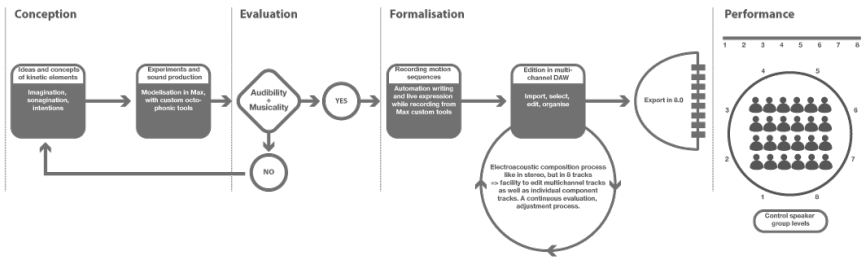


Fig. 1. This schematic offers an overview on the design / compositional process. The conception phase consists of imagining audio kinetic ideas and making sound sketches or Max models for experimenting with them. Only elements with kinetic musicality and audibility are kept for composition. Otherwise ideas are improved or abandoned. This evaluation, which also happens at further stages, is the condition of existence of kinetic music. Formalisation is the proper compositional part of the project, where sound elements are performed, recorded and organized into a music composition. During this part of the process, it is essential to facilitate modifying the sound elements.

3.1 Ideation

Planning the kinetics from the very beginning of a composition project enables us to put movement at the heart of a project, whether the kinetic plan concerns sound design or the whole composition. Adding spatialisation once the whole composition is finished, or even once sound elements are recorded, is generally too late to achieve kinetic creation consistency. Design ideation techniques exist and flourish in design thinking

literature. Some can partly be applied to art creation as well as functional design. Here are two complementary methods which can be used to sonagine[13] kinetic sound objects:
i) An inductive method which starts from an imaginary metaphor. For example, "Swings" (see Table 2.) inspired by Edgar Allan Poe's the pit and the pendulum, imagining a giant pendulum, whistling by passing near with trajectory. Here the idea is typically physical or visual.
ii) An experimental derivational process which imagines an abstract tool for moving any sound source materials. By experimenting with various sound sources and perfusing the tool, it is possible to produce lots of musical materials, out of which the most interesting can be kept for creating musical sequences.
Ideas can sometimes reveal themselves as being well sonaginated, but also other times less so because they are totally abstract and unrealistic, too visually oriented or inappropriate. In reality these two methods are often used successively or combined.

3.2 Experimentation and Sound Production in Max MSP Custom Octophonic Tools

How can a sound idea like "Swings" (Table 2) be rendered? How about synchronising between variations of timbre, position, spectral mass and spatial spread/diversity? DAWs environments are quite rigid, inconvenient for creating sounds, especially with dynamic parameters' variation when including spatialisation. Circumventing the complexity and haziness of channel management can easily be time consuming. Therefore, it is more efficient to develop fast MaxMSP sketches for each idea. These procedural models can also be improved during the experimental phase until they reach a sufficient precision to fulfil functional and expressive requests.
Two examples are described here:

- *OctoLine* allows to control a set of parameter's variations for a single stream, using break point functions to modify dynamically the following: volume, position, a filter cut-off frequency, spread and spatial diversity.[14] Octoline therefore produces variations in spatial extension. It uses mostly noise or simple generators the instances of which can be easily differentiated.
- *KDvector* launches polyphonic spatialised sound vectors triggered by MIDi notes. Time dynamic parameters or parametric vectors can be set before launching them, such as sound type volume curve, position curve, filter cut-off frequency, filter quality and spread. Diversity is ensured by multiplying vectors units.

[13] Sonagination is a neologism invented by sound designers to draw a distinction between visual imagination and auditory imagination.

[14] Spatial diversity is the desynchronization, or the originality, of the different voices. It has been used by Charles Verron [14] in his thesis to simulate natural immersive sounds such as rain. Diversity equals 0 when 8 sound voices are mixed together and played on one or more speakers according to the spread value. Diversity equals 1 when a different sound voice plays on each speaker. Between 0 and 1, positions spacing of 8 the sound voices vary continuously.

3.3 Recording Sound Elements

During the process of developing MAXMSP patches and producing kinetic shapes, sequences of octophonic sound are recorded. Each sequence is performed live or automatized to produce the desired kinetic sound. All elements are recorded in multichannel with their spatialisation embedded and will be manipulated later on in multichannel.

3.4 *Kineticality* Evaluation of Sound Elements

Before going further in the composition *kineticality*[15] is evaluated to validate sounds the kinetic expressivity of which are audible and thus to exclude those that are not[16]. A proper evaluation would require a large listener panel, but a simple listening evaluation is better than nothing. Therefore, waiting a few days before re-listening helps to step back and allows us to forget the original perception. François Bayle use to say "*il faut faire fonctionner l'oubli* "[17], in order to recover primary unbiased musical sensation. Such a subjective evaluation would not satisfy scientific evaluation criteria[18] but seems good enough for creating music for which the main evaluation is the final performance.

3.5 Editing and Assembling Pre-recorded Octophonic Elements in a Multichannel DAW[19] to Compose Musical Sequences.

Once sound elements are recorded and their *kineticality* validated, a selection of sequences or parts of them are imported in the DAW. This is where the main composition is achieved. It is important that DAWs allow the following features:

- Accept mono, stereo and multichannel audiofiles
- Edition of multichannel tracks: splitting, editing and merging. This essential feature, which allows to modify easily and precisely previously spatialised sound elements, does not exist as such in most DAWs.
  - Splitting for edition: easy switch from grouped tracks to individual tracks
  - Easy edition of the individual tracks
  - Merging edited elements: easy backwards switch from individual tracks to grouped tracks after edition
- Multichannel effects and in/out routing for multichannel effects
- In place off line multichannel effects

In Avid Pro Tools HD for example, it is easy to move multitrack clips to n mono tracks and to edit individual tracks, but in order to bring back the elements to a multitrack clip after being edited, the group must be consolidated and then moved back to its original multitrack.

[15] See Kinetic audibility §2.2, above.
[16] Nevertheless, all the sounds of a kinetic music are not necessarily kinetic.
[17] 'One should use the function of forgeting' (my translation).
[18] It should be tested by non-expert listeners using proper perceptive evaluation methods.
[19] DAW is an abbreviation of Digital Audio Workstation.

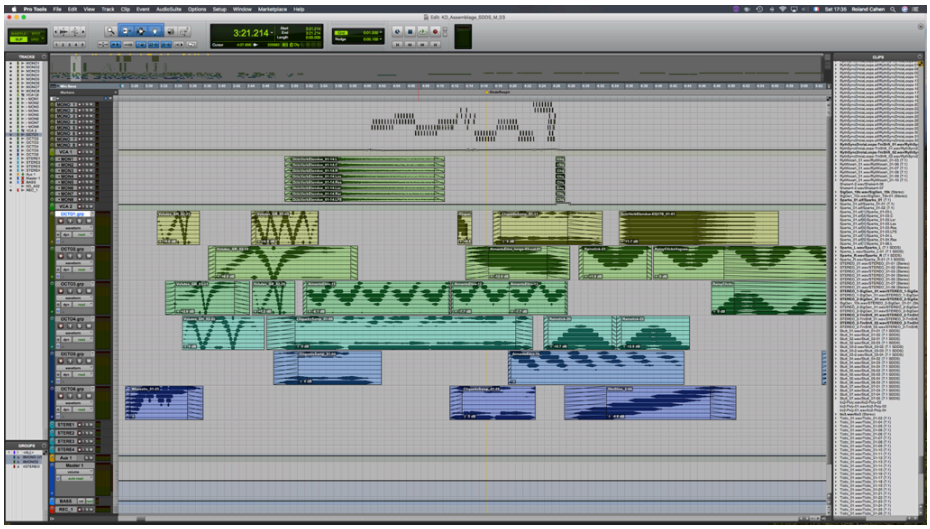


Fig. 2. Example of *Kinetic Design* Pro Tools HD Cession with kinetic multitrack elements (octophonic tracks) and mono elements in mono tracks (on top tracks)

In other professional environments, such as Steinberg Nuendo, it is possible to route any kind of track to an High Order Ambisonic bus and monitor the rendering on any number of channel outputs. But a splitting and merging process is not available unless whole tracks are duplicated for splitting again when merging.

3.6 Performance: on Groups of n Speakers and Effects on an Ensemble of Speakers (Acousmonium, Sound Dome or any Other Concert Diffusion Device)

In order to obtain the best possible kinetic perception in diffusion for all the audience, a few criteria should be taken into account:

- The acoustic of the venue should be non-resonant: muffled venues, studios or open air are preferable to empty resonant rooms or churches.
- Speakers are organised by groups of n, corresponding to the number of channels in the music or its master soundfile e.g. 8 channels => one or several groups of 8 speakers.
- Listeners are not too near the next speaker. A distance of a few meters allows to reduce the blinding effect[20]

Diffusion Setup

[20] Blinding effect in electroacoustic concerts happens when as a listener you mostly hear the speaker next to you.

The octophonic (8.0) octal diffusion format[21], which is an interesting work format because it is efficient and accessible as lots of multichannel devices work by 8, can be used. Track numbering from back left to back right allows i) to avoid an automation break in front of the listener ii) to switch and transit easily from a circular to a linear front screen distribution[22].

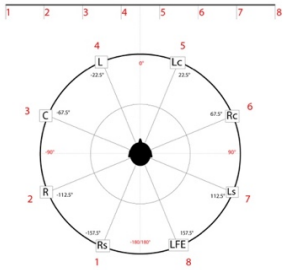


Fig. 3. Octal or circular 8.0 distribution allowing: 1- a geographic allocation corresponding to spatial intuition 2 – the possibility to switch easily from circular to linear diffusion 3 -rejects automation hedges values in the back (instead of in the front in standard formats) 4 – a simple compatibility with 7.1 SDDS standard format: L, Lc, C, Rc, R, Ls, Rs, LFE.

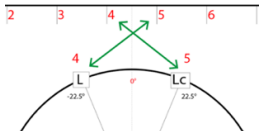


Fig. 4. Circular-linear transitions in performance

4 Excerpts of Kinetic Design Composition Logbook

5 Table 2. Kinetic Design is composed of 12 chained movements. (This table is presented here to simplify references)

	Time	Title	Duration
1	0:09	Stretching - Extension	1 ' 45"
2	1:54	Swings	2 ' 21"
3	4:15	Dots & Lines	3 ' 10"

[21] 8.0 is a common format but not a market standard, but it is compatible with 7.1 SDDS.

[22] Some parts of the music composition would avoid 1/8 transitions in order to be performed on a linear speaker array instead of a circular one. During performance, it is possible to avoid jumps when swapping from a circular to line, or line to circular, diffusion device by choosing the nearest speakers to ensure continuity e.g. 4-5. (Fig. 6.).

4	7:25	Pulsed	2 ' 54 "
5	10:19	Untied - Incises	2 ' 35 "
6	12:54	March	55 "
7	13:49	OtoGliss_1	1 ' 26 "
8	15:15	Little Soldiers	4 ' 19 "
9	19:34	Wave breathing	1 ' 06 "
10	20:40	OtoGliss_2	1 ' 47 "
11	22:27	*Laché*	2 ' 14 "
12	24:41	*Final Balancé*	1 ' 43 "

Thursday 19th July 2018

I am in the process of finishing the first part of my research to create *Kinetic Design*. My original intention was to produce a composition, or "ballet" made up of "lines" and "dots":

- *Lines between speakers: elementary trajectories, vectors or curves.*
- *Dots or punctual elements on each one of the speakers themselves, either to start and end the musical lines, to create strong accents, impulses in the music, points of support or clicking effects.*

Although the idea works, the result does not fit with what I imagined (or more accurately sonagined). To maximise spatial precision, I choose to work with two categories of sounds: impulses and sustained noise. I will refer to these figures as impulse and sustained noise spatial vector[23] (or NoiseVect).

My NoiseVect lack musicality, quickly becoming boring to the ear, almost like a dancer stretching her arms back and forth. Worse, they do not validate the idea of a line, segment or trace, especially when they are slow. Faster motion and shorter swishes, such as mechanical pistons or whip sounds give a better idea of a directional movement. But I could not find a way of making them sound like lines starting from a position and finishing on another. And if we can hear whipping sounds crossing through space, the sounds also carry a dramatic and ridiculous musical connotation. It even seems that line/trace impression are inversely proportional to spatial precision. In other words, the more accurate a position, the more noise the sound must carry. However, noise is perceived as large and blurred and not as a line. On the contrary, sustained sounds, reduced in mass (FN), such as pure tones, which better suggest lines, can hardly be heard at a precise position in space. This leads me to having to choose between Scylla and Charybdis.

[23] Vectors are elementary directional trajectories going from one speaker to another.

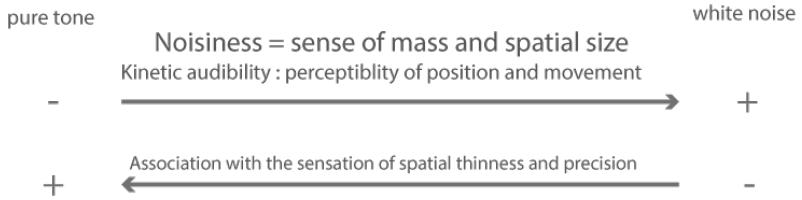


Fig. 5. Impression of a trace thinness being inversely proportional to spatial resolution. Impulses placed at a precise position on one single speaker (track) are situated well in their position, but neither give the idea of points. They sound rather like events marked in time and space. When an impulse is attached to a *NoiseVect*, at the begining or end, it sometimes appears grouped together, sometimes not[24]. But even when impulses and *NoiseVect* merge together in time and timbre, impulses are not perceived as being beginning and ending positions of a Noise*Vect*, since the positions do not seem to merge. Spatial perception thus seems to keep separate from the sound stream. Further research could be done on merging spatial and sound content perceptions. Fortunately, as soon as rhythm and musicality inhabit these sounds, I enter as a listener into an astonishing and new sound universe. Like a sorcerer's apprentice, I feel incapable of understanding exactly what I manipulate. I can hear some effects, artefacts, bursts, spreads, folding and unfolding, but I cannot manage any of these elements precisely. While writing this, I am reminded of Pierre Schaeffer, whose *journal de la musique concrete* [10] has been an inspiration for me.

Monday 27th August

Capping/unplugging effect: difficult to achieve with simple clicks, try breaking an infra-bass hold sound with a strong distributed click. Try curtain effects from mass of scattered grains passing from one side to the other opening and closing a background.

– MAKE SILENCE, rarefy, nuance, intensify differences.

Mobility takes its value only in contrast with stationary moments. Look for motion variation control.

– For the swings: hold the balance in a position creating tension, then release. Try Ping-Pong rhythms on the origin and the destination of the pendulum (Table 2).

[24] As established by Albert Bregman's grouping principles [4].

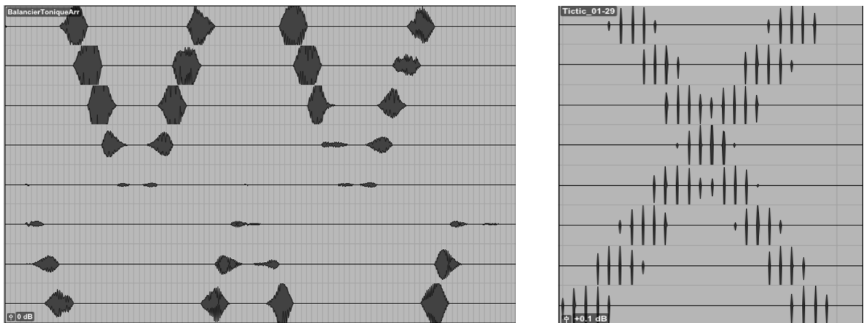


Fig. 6. Transitions (Left) Rear swing: transition from 8 to 1 happens in the middle of the movement - (Right) Two identical iterative sounds in opposite direction, starting from the rear, crossing forward and ending at the back gives this amusing picture (1 to 8 + 8 to 1)

Note that the intersection of the two iterative sound movements is not in the centre but on channel 4. If the crossing happened in the middle, the sound would be played on speakers 4 and 5 with a bad spatial focus, and would reduce the spatial resolution to only 7 points. With 8 channels this kind of asymmetry is inevitable, but on the other hand, odd numbers of channels would bring other asymmetries, e.g. rhythmical ones. Twelve channels would be a very good working module but unfortunately not standard and more expensive.

Tuesday 28th August

The GRM studio-A is comfortable and beautifully equipped[25]. The acoustics are excellent, the sound awesome, the computer powerful. A shelter from the outside world in a hidden basement, where there are no distractions. Perfect!

I have tried many musical concepts, recording hours of octophonic sequences using different models: plots, whips, swings, spatial loops, variation of position, diversity, polyphony and spatial amplitude, with various sound materials: dynamically filtered white noise, various wave generators and samples. On the whole there are some remarkable figures, but I am struggling. Once these sound movementsare produced, it is difficult to assemble them. I tend to regress, replaying musical approaches of the late 70's called "séquence jeu", where the same set of gestural sounds are repeated with pattern variations, quickly becoming monotonous. Trying other formal approaches, I combine different parts of the composition as if they were sediments, layers of a landscape, as if each one of the parts of the work were stacked on top of one another. I play with transitions, creating geological faults in the landscape, planes and protrusions, leaving glimpses into the future or flashbacks into the past.

I realize that my sampled loops are spatially shifting because of a scheduling drift between the signal and data, which I then correct. At last, each sound element in a sample

[25] The studio is equipped with 8 Genelec 8250 monitors placed around the central working position.

or loop now plays in sync and always comes through on the right speaker allowing me to make kinetic rhythms. I can also precisely control the position of each component.
I have created my own instruments with Max, which gives me a lot of possibilities, but the difficulty with Max is that I develop and add new features all the time, and the patches are never stabilized. Each time I try to add a new feature to the patch, new bugs appear and I need to debug from time to time. Moreover, it is impossible to memorize the best presets because after each change the settings previously memorized become obsolete. Consequently, I record audio at each step and use Max "pattr" presets only as setting-up facilities.
Kinetic masking effect and kinetic pattern segregation - the masking effect[26] acts at the spatial level. It is possible to play on the contrast of high-pitched moving sounds mixed with sustained bass sounds, or on close planes in relation to more distant planes. But as soon as two simultaneous movements or positions are exceeded, independent movements become quite difficult to hear, except with very different or complementary or eccentric sounds. Hence the use of very eccentric registers, e.g. the infra-bass only slightly disturb the spatial perception of other sounds. Maximising temporal, spectral, dynamic and of course spatial differences helps to separate the patterns.
Playing with the diversity parameter - diversified sound masses, i.e. composed of 8 independent sound sources, but of the same nature (e. g. 8 different white noises on each track versus the same white noise on the 8 tracks), allow me to stabilize the movements of sounds and give our ears a rest. They create a sensation of breadth and richness, like open-air soundscapes.
MAKE CLEARER - spatial complexity and diversity must be undressed, flattened or lightened, the movements should stop. More static and monophonic moments would reinforce the spatialised ones. Should I do this in the composition or during the performance? Both probably. Unfortunately, the possibilities to reduce complexity without losing content during the performance are limited.

Thursday 27th December
I am at the end of the project. The composition is more Baroque and more dramatic than originally expected. Many bass and infra-bass sounds, clinging clicks and stridencies provoking inner ear interferences. Treble strident sounds are heard as they saturate the ear itself. Strange feeling, It seems as if the sound was produced by the ear itself, often mixed up with provoked otoacoustic emission, I do not think they are the same phenomenon. They happen when treble sine which are distributed in frequencies and space and are modulated slowly, typically between 1 and 4 kHz. At the end of listening, the ear is physically tired, yet the listener has travelled to unusual places and sensations.
I believe I have succeeded in my challenge to create kinetic music, where spatialisation plays a central role, constitutive of the musicality. There is a lot of imagination and variety in kinetic configurations allowing new sounds and musicality to emerge. There is hopefully more to explore in order to create music with new sounds, finding their

[26] Auditory masking effect occurs when the perception of A is affected by the presence of sound B. It also concerns simultaneity, frequency, spectrum and directional masking

own expressiveness. There is also humour in there, like the grotesque walk of the little soldiers, both burlesque and tragic, ridiculous and terrifying at the same time.
It is difficult to find the right balance between movement and immobility. Like watching a dancer on stage, and appreciating the choreography, it is easier if the stage and the spectator do not move too much themselves. Here there are sometimes too many unnecessary swirls, but I have a hard time deciding which ones to cut in order to stop the movement in the right place, moment and positions. Two approaches are possible: either :- i) starting from fixed sounds and building the movement at the time of the compositional process, which means adding a disconnected movement on top of an existing playing sample ii) or pre-constructing the movement inside the sound materials themselves and then reducing or fixing them afterwards, with the inconvenience that the movement is fixed and attached to the sound. I mostly used the second method.
I have moved on from the idea that sounds have to be noisy or contain noisy attacks to be precisely positioned, while tonic and sustained sounds are difficult to locate. That is still true, but intermediaries also work. It is an important compositional choice that the positions be clear and precise for some sounds, less for others. Trajectories require a fairly high spatial precision, unless listeners simply feel that spatialisation has changed, but without perceiving kinetic qualia. Spectral, temporal, pattern complementarity is also a major tool for distinguishing kinetic qualia. For example, impulses are very easily positioned on top of sustained sounds and seem to appear at the forefront. This complementarity is a key to access musical complexity.
Kinetic design is a composition that must be listened loud enough but not to the point of needing ear protectors. Physiologically speaking, it is very demanding for the listener's ears. At the end of listening to it, listeners may start to feel this composition has made them work in various unusual registers but without pain or hearing fatigue.

6 Conclusion

This design process has proved efficient in a number of ways. A large part of the audience who listened to *Kinetic Design* said they had a singular experience of spatialisation. For the first time, they could really hear the movement of the sound. The spatialisation was effective and impressive. They also said that it was a show of sound, the plasticity of the sound was expressive and constituted a novel listening experience. However, does this mean that there is a new musicality, a kinetic music, a new genre in music? The answer is more nuanced. Some other composers working in the same field who listened to the performance were quite critical and said that the music did move but they could not feel any particular significance in the movement, nor a specific kinetic musicality. Any evaluation from listeners' experience is only representative of a sample; it is bound to be subjective and must therefore be relativised and treated with caution. The question whether kinetic music can be developed into a new expressive genre or is only a fictive construction to push spatialized expressivity forward to its limits, still remains of interest and needs to be explored further with more

experimentation, applied research and experimental creation. New tools and design processes also may help to improve kinetic music experience.

References

1. Baalman, M.: M.A.J. 2008. On Wave Field Synthesis and Electro-Acoustic Music, with a Particular Focus on the Reproduction of Arbitrarily Shaped Sound Sources. PhDthesis, Technische Universität Berlin
2. Baalman, M.: Spatial Composition Techniques and Sound Spatialisation Technologies, Organised Sound15(3): 209–218 & Cambridge University Press, 2010
3. Blauert, J.: 1997.Spatial Hearing. Cambridge, MA: TheMIT Press.
4. Bregman, A.: A.S. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, Mass.: Bradford Books, MIT Press (1990).
5. Cahen, R. : Octophonie DeltaP-Cahen. Compte rendu de recherche pour Octophonie Studio Delta P La Rochelle (1994).
6. Hauberg, J.: Research by Design – a research strategy
7. Macedo, F.: Phenomenology, spatial music and the composer: prelude to a phenomenology of space in electroacoustic music: Proceedings of the International Computer Music Conference 2011, University of Huddersfield, UK, 31 July - 5 August
8. Merlier, B.: Vocabulaire de l'espace en musiques électroacoustiques DELATOUR France, pp.230, 2006. halshs-0051174
9. Pottier, L. La Spatialisation Des Musiques Électroacoustiques. Université Jean Monnet-Saint-Etienne. Edited by C.I.E.R.E.C Travaux ; Musique et Musicologie 157. Saint-Étienne: Publications de l'Université de Saint-Étienne, 2012.
10. Schaeffer, P. : De l'expérience musicale à l'expérience humaine. La Revue musicale n°274-275. Paris, Richard-Masse (1971).
11. Schaeffer, P. : Traité des objets musicaux essai interdisciplines. Paris: Éditions du Seuil, 1977.
12. Schaeffer, P.: Machines à communiquer Tome 1 Genèse des simulacres
13. Thigpen, B.: Spatialization Without Panning. eContact! 11.4 Symposisum électroacoustique de Toronto CEC https://econtact.ca/11_4/thigpen_spatialization.html , last accessed 2019/04/20 (2009).
14. Verron, C. : Synthèse immersive de sons d'environnement. Thèse de doctorat en acoustique Université Aix-Marseille (2010).

Augmented Live Music Performance using Mixed Reality and Emotion Feedback

Rod Selfridge and Mathieu Barthet *

Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London, United Kingdom
{r.selfridge,m.barthet}@qmul.ac.uk

Abstract. This paper presents an experimental study into the use of mixed reality (MR) visuals generated from performers' expression and emotion sensing augmenting a live music performance. A laptop instrument was used to trigger melodic samples which influenced MR visuals based on musical properties. The MR visuals were viewed either on a Microsoft HoloLens or a mobile device with participants free to move in the performance venue as they chose. An emotion sensor tracked participants' facial expressions and predicted their expressed emotion which was mapped to emoticons on a screen. Results show that the MR visuals were very positively received but less so the emoticons. Results also provide guidelines to improve such system by adapting the visuals to better fit the field of view, supporting mapping of more subtle expressive variations, and interactive content that can be curated.

Keywords: Mixed Reality, Augmented Reality, Emotion Sensing, Multimodal Live Music, Internet of Musical Things, Smart Musical Instruments, Microsoft HoloLens

1 Introduction

Western live music performance settings have been relatively static since the advent of electroacoustic instruments and amplification. Productions of popular, electronic and experimental music have embraced live visuals and lightning effects accompanying the performance, leading to spectacular shows, however the visual content or effects are often based on pre-conceived material synchronised by video Jockey (VJs) and lightning engineers [1]. Live audiovisual systems leveraging performers' expressions as performed on stage in real-time are scarce. This poses challenging signal processing and ICT problems as performative expressions are difficult to capture and analyse due to their intangible nature and the lack of representations at the symbolic level (e.g. score). Furthermore, to date, the augmentation of live music performance through other sensory modalities is performer-centric and does not take into account the audience's response, which is fundamental for the completion of the artwork and the overall experience. This work is part of a larger project examining how mixed reality (MR) can support novel aesthetic narratives driven by both performers and audience's expressions over the course of live performances. MR head-mounted displays (HMDs) enable

* This work was supported by Innovate UK project "ALIVEmusic - Augmented Live music performance using Immersive Visualisation and Emotion".

to "keep the eyes and the feet" into reality while experiencing computer-generated audiovisual content. As such they appear very relevant to social context such as live performance, during which virtual content can complement or blend with the musical and embodied expressions performed on stage.

In Western live music performance, there is often an unspoken expected audience etiquette, which despite great differences across musical styles (from sitting quietly and applauding at the end in classical music, to dancing and yelling of excitement in rock concerts), confines the audience to a role of receiver, in a creative sense at least. Although very powerful and engaging live musical experiences have and still are delivered, both scholars and artists have questioned the definition of the concert setting, models of musical communication [3], and reflected on how to improve audience engagement [2]. These reflections are timely in an era where the music industry is shifting with music media and their delivery being reinvented, placing more importance on interactivity and creative agency [4, 5]. Audience participation has been proposed as a means to reduce the fourth wall between performers and audience [3], facilitated using emerging technology-mediated audience participation tools [4]. In this study, we start to investigate the proposal that MR could be used to reduce the divide between performers and audiences by offering a new medium of communication through which performer-audience associations and bonds could be realised.

Audience engagement (and therefore participation to live concerts) can also be restricted by a lack of understanding of the music, context, or expression of the performers. For example, electronic instruments often require the performer's attention to be focussed on tangible and screen-based interfaces [6]. This disconnect between gesture and the physical sound producing mechanisms make it more challenging for audiences to comprehend performers' actions when on stage. It is stated in [7] that when performing, gestures can express skill, control and introduce an aesthetic component which can be lost by the use of electronic instruments. In this study we start to investigate whether the use of MR can assist an audience's understanding of performers' expression. In the remainder, after discussing related work, we present our study and findings.

2 Background and Related Work

Research into musical experiences within an extended reality (virtual, mixed and augmented realities) environment is a growing area. Virtual Reality (VR) has seen a large growth in research across a wide range of disciplines shown in [8] and music is one of these. In [9], the author examines the potential of virtual reality in providing new expressive musical experiences (*"visceral, aesthetic response to the music in a way that was connected to the immersive visual environment"*). He concludes that there are no real experts in the field of VR for music and understanding is continually developing. In [10], the use of an interactive music system within VR allows users to create, experience and engage with music in a new and novel manner. Findings indicated that new users enjoyed interacting with music within the virtual environment although confusion on how to interact with the system as well as the mechanism for controlling virtual objects could often be prohibitive and frustrating. Using musical instruments

within a virtual environment is presented in [12], exploring novel methods of control and interaction. It is identified that new immersive environments offer musical experiences beyond those of performing with traditional instrument and indicates that an intelligent visual feedback system may aid performance.

The Internet of musical things (IoMusT) proposed in [13] and [14] describes the networking of devices into musical objects by use of embedded computing devices, for the purpose of producing or receiving musical content. Unlike the aforementioned virtual musical experiences, the musical things are generally real world objects which can connect to other real or virtual objects. The associated ecosystem supports novel interactions between musicians and audience members, impacting on how music is created and experienced. Within this ecosystem, smart musical instruments are instruments which can include embedded sensors, actuators as well as wireless connectivity [15]. The augmentation introduces novel pathways for expression that are often not possible on the same instruments without additional sensors, etc. Similarly, sensor-based gestures can lack a large sound variation and novel playing techniques are often developed to enhance and incorporate any limitations [15].

A 3D visualisation was provided in [16] which augmented an electronic instrument. This study identified the disconnect between the performer of an electronic instrument and the audience members due to limited visible gesture, giving the example that an laptop performer may have just pressed play on a pre-recorded track and is checking their email through the performance. This study provides a 3D representation of sound processes of the instrument linking the sensors to a visualisation, revealing how the performer is impacting the sound. A similar study where 3D visualisation augments a musical performance is presented in [17]. In this study a seated audience view a performer interact with 3D virtual objects affecting the graphics and sounds. The audience were directly able to hear the performers musical expression and see a visual representation. One issue was that due to the projection of the visuals, the performer had a far different perspective than the audience members.

Augmentation of a performing stage was carried out in [18] where smart phones were used by audience members to view an augmented environment. Augmented reality (AR) elements are placed on the stage which can be viewed on the audience members smartphone. Once the smartphone has detected the AR object, the audience member can manipulate these and contribute to the performance outcome. A significant finding from this study was that at least one performer stated that the would like additional rehearsals with selected audience members. It could be argued that if audience members are attending rehearsals to coordinate the musical outcome, she/he would no longer part of the audience and have become performers in their own right.

The importance of the emotional expression of music, perceived and felt, is well known [19]. [11] proposed a music visualisation system augmenting the listening experience by responding to both the music and the listener's arousal as characterised by the electrodermal activity (EDA). The density of the visuals are influenced by EDA levels which tend to increase during excitement or stress

due to the skin's perspiration. The present study focuses on the use of a MR environment where the audience and performers can both see holograms (3-dimensional virtual projections) and additional visuals reacting in real-time to the music being performed and the audience's emotional response.

3 Experimental Setup and Method

3.1 Hardware and Mapping

MR visuals were delivered via a Microsoft HoloLens HMD along with augmented reality experienced through use of an iPad and iPhone. These technologies give participants freedom of movement, ability to interact with each other and maintain a link between performers' actions and the music. The electronic instrument on this occasion was a laptop, used to trigger sample sounds. Eight individual samples of a percusive hand pan drum were mapped to keys on the laptop, each triggering once when pressing a key. The 8 samples were the notes F, G, Ab, C, Db, F, G, F and C, corresponding to an F minor scale. As highlighted in previous studies [9, 12–14], ensuring low latency between the hardware is important. Data communication between the laptop and the MR displays was achieved by sending messages adhering to the OSC protocol via a WiFi router [20]. An emotion sensor comprising of a video camera tracking participants facial expression was also placed within the space which enabled participants emotion to be indicated. The use of these technologies allows for MR visuals to be generated in real-time by a performer, giving them an additional medium for aesthetic expression. As a form of a visual feedback system we indicate the participants' felt mood by way of an emoticon projected onto a screen, visible to both audience and performers. The hardware data flow and feedback via participants is shown in Fig. 1. The audience experience the music as both sound and vision, their emotional response gauged by the emotion sensor. The applications running on the HoloLens and mobile device were designed to augment the performance with a hologram projected in the space. The hologram was a sphere of coloured balls whose shape and position reacted to the music. Mapping between the audio and the visuals was achieved via non-disclosed proprietary software from our industry partners, Fracture Reality[1]. The visual display of the mobile device can be seen in Fig. 2. For this experiment, the audio was played from the MR devices own speakers. The emotion sensor used to predict the emotion of the participants throughout the experiment was from our partners Sensing Feeling [2] and analysed facial expressions from a webcam's video stream. The system is based on deep learning computer vision techniques described in [21]. The sensor outputs data relating to a *customer delight index*, a measure of satisfaction within the sensor range (from low to high), measured from facial expression also factoring in the number of people. The output from the sensor was processed in real-time and emoticons relating to each individual in the sensors view projected onto a screen facing the audience and on the side of the performer to provide feedback. The background in Fig. 2 shows two emoticons relating to the participants. In this instance the emoticons only had two states;

[1] http://fracturereality.io/
[2] https://sensingfeeling.io/

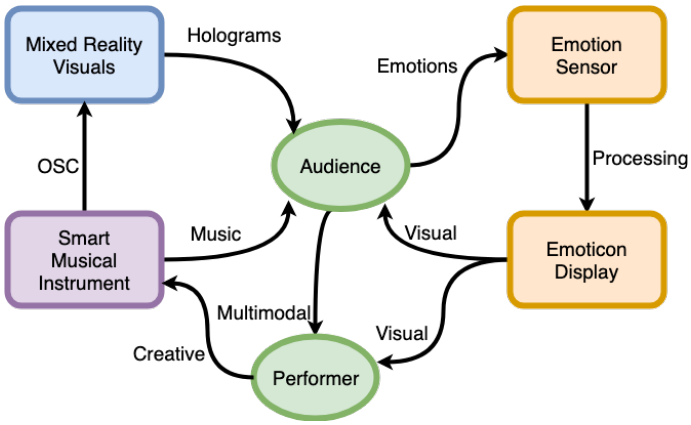


Fig. 1: Data flow between hardware, audience members and performers

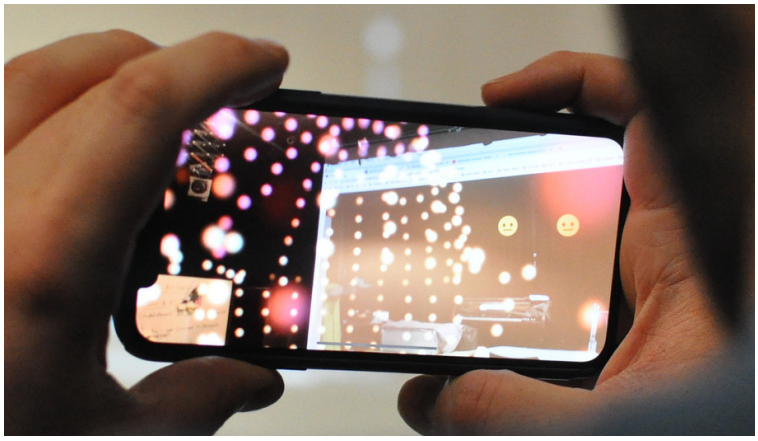


Fig. 2: Music-driven mixed reality visuals as seen through a mobile device

a smiling face and a neutral face. Processing of the emotion sensor occurred online, returning emoticons based on the expression of audience members which were projected on a screen at the back of the room. The emoticons provided a feedback to the audience and performer of the emotional response of the audience members. For the performer, this provided an additional form of feedback from the audience along with the more traditional multimodal feedback methods - body movement, clapping, etc.

The musical pieces were performed on one smart instrument (laptop) by one of the authors, lasting 2-3 minutes in length to allow all participants to experience the MR visuals through the HoloLens and the mobile device. The music was an improvisation with a relatively constant tempo and similar pattern over a pentatonic scale played by performers across the test groups. Participants were asked to give their perspectives as audience members or performers according to their musical experience. In this test, all participants experienced the experiment

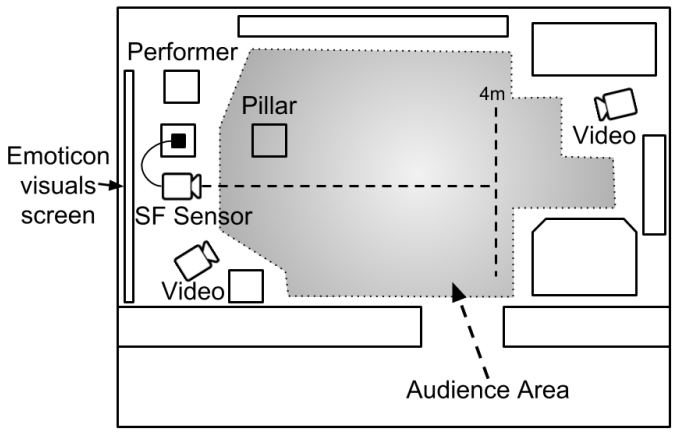


Fig. 3: Room setup

as an audience member but feedback obtained was tailored depending on the participants potential use of the system, (audience member or performer). Due to the nature of the current laptop instrument, there was not a wide range of musical dynamics available for artistic expression.

3.2 Room Setup

The tests were carried out in the Performance Lab, Queen Mary University of London, with the room setup shown in Fig. 3. The emotion sensor's optimal distance from participants is 4 metres. Participants were able to move anywhere within the room that didn't interfere with the experiment equipment. This is shown as a shaded area in Fig. 3.

3.3 Methods

To collect feedback from the participants they were asked to complete a questionnaire at the conclusion of the performances. The use of MR visuals with emotional feedback as is classified as an expressive interface within the taxonomy relieved in [7]. Interfaces classified as expressive can reveal or amplify manipulations to enable participants to appreciate how a performer is interacting with the system. The questionnaire was constructed to identify if participants perception of the performers musical expression was enhanced by use of the proposed system.

Feedback was gathered through a mixture of responses from participants. The first questions were free text, directly enquiring if participants enjoyed their experience as well as particular aspects they liked and disliked. Following this a series of statements were proposed, requesting participants responded via a Likert items ranging from 1 to 7; 1 representing that they strongly disagreed with the statement and 7 representing that they strongly agreed with the statement. For each of the Likert questions, participants were given the opportunity to give a detailed explanation as to their choice.

3.4 Participants

A total of 11 unpaid audience and performer participants took part in this study, with a total of 5 tests run in total; 2 participants in 4 of them and 3 participants

in 1. 5 participants gave their perspective as audience members, (1 female, 4 males) with a median age of 29. They all had experience with VR, 3 with AR, 1 with MR and 3 with emotion sensing. They attended 1 musical performance per month. 6 participants gave their perspective as performers, (1 female, 5 males) with a median age of 30. 4 had experience with VR, 4 with AR and 2 with MR; the emotion sensor was new to all performer participants. They performed 1 musical performance per month.

4 Results and Discussion

4.1 Quantitative Analysis

An important aspect of this study was to investigate if participants felt more connected to musical expression following the augmented performance. Musical expression is associated with properties such as tempo, interval, rhythm, articulation, etc. [22]. It is through musical expression that composers and performers convey the emotional content, which are either felt or perceived by the audience [19, 22]. Discrete distributions illustrated give a more detailed breakdown of preferences compared to mean with interquartile range.

It can be seen from Fig. 4 that the majority of participants felt more connected to the music than usual, especially performers, yet it could be argued that, when expressing if they were able to perceive the musical expression more clearly, two groups are present, Fig. 5. For audience members, it is hypothesised that those who perceived the musical expression more clearly were less distracted by the novel technology and saw the visuals as a complement to the music while others may have been distracted by the visuals. A negative response was given in relation to the emoticon visuals, where all participants disagreed that the emoticon visuals influenced their appreciation of the music, Fig. 6.

Figure 7 shows the responses when participants were asked if they would like to view the emotional response of the other group, i.e. audience viewing the performers' emotional response and performers viewing the audience response. Results indicate that there seems to be two separate trends in the group. One set of each group participants do not wish to visualise the others' emotional responses but clearly another group do wish to see them.

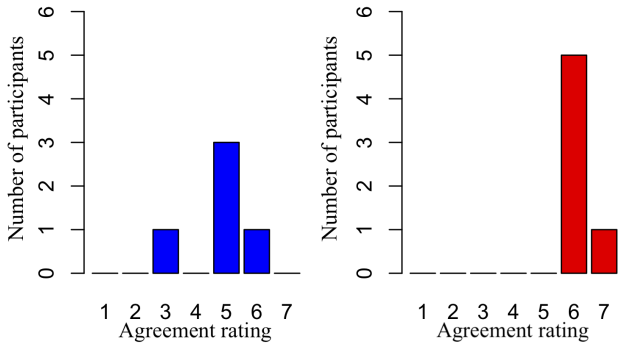
4.2 Qualitative Analysis

Along with giving feedback via Likert items, participants were asked to comment on their choice of response, including free text on some open-ended questions. An inductive thematic analysis in accordance with [23] was carried out to analyse these responses and a number of themes were identified from the responses, as indicated below (number of codes in brackets).

Audience: Constraints (19) The field of view of the HoloLens was by far the greatest complaint from participants (*"The limited viewpoint of the headset stopped me feeling fully immersed"*). The lack of expression from the laptop instrument was also felt to constrain the performance (*"the instrument does not allow much expressiveness"*) and holding a mobile device *"felt clumsy"*. **Expressiveness (15)** The link between the music and MR visuals was clearly felt by a number of participants (*"music drove the visuals"*). Some of the audience

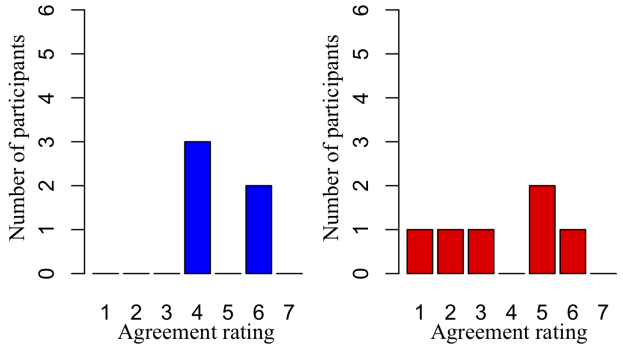
members felt the visual augmentation gave them a clear sense of the performers' actions (*"The closeness of the virtual object made me feel closer to the performer and their musical expression"*). **Participation (14)** The number of codes for participation from audience members indicate that this was high as a priority for them. Building on passively viewing the visuals making some feel closer to the performer, influence on the performance is also desirable (*"I could play with the visuals myself, disrupting their connection to the music in a playful way"*). One participant also indicated that having their emotional response would be *"a nice feedback mechanism"*. **Engagement (13)** Audience participants have indicated that they found the system more engaging as a whole than if the music played during the study (simple hand pan drum melodies) had been played alone (*"Visuals extend my perception of music and enrich my experience"*, *"kept me focused on the music"*). **Reservations (13)** A large number of reservations were raised in relation to the system, a number of these were in relation to a distinctive representation of the individuals' emotional response (*"I wouldn't want to visualise how I feel"*). Other reservations related to whether adding technology to a music performance adds value. **Disregarded (12)** Although there were a number of codes on this theme, they were focused on one issue, the emoticons which were often disregarded (*"I occasionally looked at the emoticons but they never changed so I basically ignored them"*). **Improvements (11)** The number of improvements suggested by audience members all focused on changes or enhancements to the MR visuals and their connection to the music (e.g. *"colours instead of shapes"*, *"multiple virtual objects scatter through space"*). **Social Environment (8)** MR offers participants the opportunity to share some aspects of the experience with others (*"I really like that I am in the same room and see the people that are inside of it rather than being catapulted in to a whole different dimension like VR"*). One audience participant felt that they *"experienced this independently of others"*, while another noted *"I could feel comfortable in my surroundings whilst being immersed in the performance"*. **Enjoyment (4)** Although enjoyment rated high in the quantitative study, it was directly quoted in the free text only to a small extent. **Distraction (3)** Two of the comments here stated the participant was focused on the visuals which *"distracted me from listening to the music"*, while another participant found the latency between the different devices a distraction.

Performers: **Improvements (29)** Not unsurprising, performer participants had a number of ideas on how they would like the system to improve and how they might incorporate it into their own performances. The majority of comments were in relation to the connection to the visuals and the expressiveness of the musical instrument (*"If the system were able to respond to legato, intensity, projection, tessitura, warmth etc., that would be amazing!"*). One stated that they would like it if *"the audience could see a visual representation of the feelings that I am trying to convey"*. Another stated that they would like the MR visuals to *"display aspects of my software and equipment"* and giving them *"virtual controls"* for their equipment. **Reservations (22)** The majority of reservations expressed by performers was in relation to a visual depiction of their current



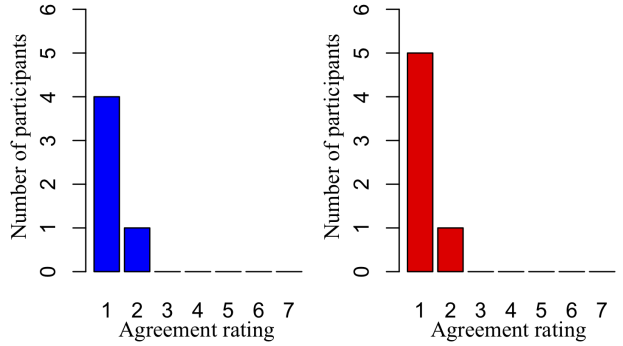
(a) Audience (b) Performers

Fig. 4: Responses to statement "I felt more connected to the musical expression than usual"



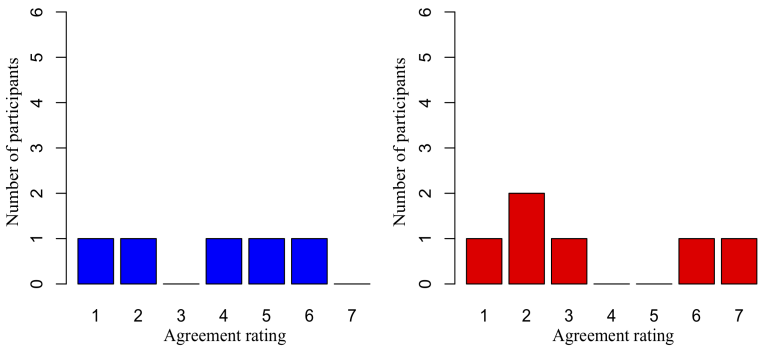
(a) Audience (b) Performers

Fig. 5: Responses to statement "I was able to perceive the musical expression more clearly than usual"



(a) Audience (b) Performers

Fig. 6: Responses to statement "The emoticon visuals influenced my appreciation of the music"



(a) Audience (b) Performers

Fig. 7: Responses to statement "I would like the mixed reality visuals to represent the emotional response of the performers (left) / audience (right)"

emotional state (*"As a performer, I don't necessarily want the audience to see my true emotional state"*). Another commented that they would prefer the emotional feedback to be *"integrated much more carefully and abstractly"*. Feedback also questioned how nuanced the visuals were at revealing the intended expression (*"I didn't really connect the dynamic changes in the visuals with the music"*). **Constraints (19)** The expression of constraint of the system generally fell into two groups - the limited range of view of the HoloLens and the lack of expressive capability of the laptop instrument (*"if the instrument was more expressive this would be easier to gauge"*, *"a more velocity-sensitive instrument would allow for greater expression"*). Performers also were not positive about holding a mobile device during a performance. **Engagement (17)** Performers generally found they were engaged with the system as a whole. One comment - *"having the visualisation of the music being made gave me a clearer reflection on the performance of the musician"* - clearly indicates they connected with the augmented performance. Another participant stated they were *"thinking about future performance possibilities"* in relation to their own performances. **Disregarded (13)** All the codes for this theme related to the screen projection of the emoticons. **Participation (13)** It was found that participation was enhanced by the MR visuals (*"Breaking the separation between audience and performer"*, *"Having a more holistic experience"*). Another fascinating comment from one performer was *"I would love for this technology to have the ability to use the performer as a 'conductor' of the visuals not only musically but emotionally as well"*, indicating the potential value added by emotional feedback. **Expressiveness (12)** The majority of comments in relation to expressiveness of the system was that participants could connect the visuals and music (*"intensity and complexity of the music was enhanced by the visuals"*). One participant was more negative about the connection between the visuals and musical expression (*"musical expression was very much lost"*, *"visuals were obviously linked to the music but not really the expression"*). **Enjoyment (11)** The comments on enjoyment were focused on the ability to move around the MR visuals, the

expressiveness of the visuals and the integration with the HoloLens. **Social Environment (11)** From the comments given, a number of performers saw the system as a tool for connecting the audience members to share the holistic experience (*"It would create a sense of unity and connection both within the audience themselves"* and *"as a performer it would be important to share the same experience with the audience"*). **Distraction (9)** A number of distractions were highlighted by performers. One participant was distracted by the emotion sensor (*"overthinking how it was analysing my reaction, which detracted a bit from the immersive nature of the expreience"*). Two felt the MR visuals distracted them from the music but one thought this was *"in a good way"*.

It is accepted that more participants would have been preferred however it is argued that quantitative analysis and thematic analysis yields commonalities and differences between participants useful to inform the next design steps. Results indicate that the majority of participants appreciated the expressive connection between the MR visuals and the performance. The majority of negative comments were in relation to the emoticon visuals and the HoloLens' field of view. We found that individuals do not want to see a direct representation of their current emotional state. It is believed a more abstract representation would be more appropriate as emotional response was still seen as a valid feedback mechanism by some. Some audience members were open to the idea of interacting with the visuals but notably, none expressed a desire for this to affect the music, leaving this medium to the performer. Performers generally indicated they would like the musical instrument and MR visuals to be more expressive.

5 Conclusions and Future Developments

We presented a novel interactive system which maps performative musical expressions and emotional response to computer-generated MR content in real-time. A study was conducted in a controlled lab setting with 11 participants who used two interfaces (Microsoft HoloLens, mobile device) during short performances. Results from quantitative and qualitative analyses highlighted that participants engaged very positively with the immersive MR visuals. Initial evidence was found supporting the proposal that MR can enhance the sense of participation and connection to the musical expression, while maintaining a favourable social environment during live music. Contrastingly, participants disregarded screen-projected emoticons which were felt as intrusive and did not blend well with the performance. Visual feedback about emotion perception was seen as an interesting idea by multiple participants but preference was noted for more abstract presentation and/or one operating at the collective level. Suggestions were also made to improve the mapping between musical expression to MR visuals by taking into account a wider range of musical attributes such as dynamics, timbre, articulation and tessitura. The HoloLens was more popular than mobile devices but the field of view's restriction induced frustration. This was out with our control but will be rectified with the next generation of this technology. In future work, we plan a study in a 'gig'-like atmosphere closer to a real-world setting with sound played from a PA system rather from MR devices. Using a PA for the audio would eliminate audio latency due to WiFi communication but

might introduce a visual latency. More detailed feedback could be obtained by giving performers the opportunity to play smart musical instruments, allowing for their own musical expression to control the MR visuals, and by providing them a physical controller to curate the visuals.

6 Acknowledgements

This work was supported by the Innovate UK Audience of the Future Design Foundation project "Augmented Live music performance using Immersive Visualisation and Emotion" (ALIVEmusic, Project no. 133749). We warmly thank our partners Mark Knowles-Lee (Fracture Reality), Jag Minhas (Sensing Feeling), and their teams, for the discussions about this work and their support.

References

1. Olowe, I and Grierson, M and Barthet, M: User Requirements for Live Sound Visualization System Using Multitrack Audio. Proc. Audio Mostly. 40, 40:1–40:8 (2017)
2. Dobson, M. C. and Sloboda, J.: Staying Behind: Explorations in Post-performance Musician-Audience Dialogue, In Coughing and Clapping: Investigating Audience Experience, Ashgate, 1:1–17 (2014)
3. Kattwinkel S, editor. Audience participation: Essays on inclusion in performance. Greenwood Publishing Group (2003)
4. Wu, Y and L. Zhang, L and N. Bryan-Kinns, N and Barthet, M: Open Symphony: Creative Participation for Audiences of Live Music Performances, in IEEE MultiMedia, 24(1), 48–62, (2017)
5. Barthet M, Fazekas G, Allik A, Thalmann F, B Sandler M. From interactive to adaptive mood-based music listening experiences in social or personal contexts. Journal of the Audio Engineering Society. 64(9):673-82 (2016)
6. Turchet, L and Barthet, M: Co-Design of Musical Haptic Wearables for Electronic Music Performer's Communication, in IEEE Transactions on Human-Machine Systems, 49(2),183–193 (2019)
7. Reeves, S and Benford, S and O'Malley, C and Fraser M: Designing the spectator experience. Proc. of the SIGCHI Conference on Human Factors in Computing Systems, 741–750 (2005)
8. Milgram, P, and Fumio K: A taxonomy of mixed reality visual displays. IEICE TRANSACTIONS on Information and Systems 77.12, 1321–1329 (1994)
9. Wang, G: Thoughts on Virtual Reality Design for Musical Expression. ACM CHI Music and HCI Workshop, San Jose. (2016)
10. Deacon, T and Stockman, T and Barthet, M: User experience in an interactive music virtual reality system: an exploratory study. In Proc. CMMR. 192–216 (2016)
11. Subramaniam A, Barthet M. Mood Visualiser: Augmented Music Visualisation Gauging Audience Arousal. In Proc. Audio Mostly (p. 5). ACM. (2017)
12. Serafin, S and Erkut, C and Kojs, J and Nilsson, NC and Nordahl, R: Virtual reality musical instruments: State of the art, design principles, and future directions. Computer Music Journal 40(3) 22–40 (2016)
13. Turchet, L and Barthet, M: Ubiquitous musical activities with smart musical instruments. Proc. of the Eighth Workshop on Ubiquitous Music. (2018)
14. Turchet, L and Fischione, C and Essl, G and Keller, D and Barthet, M: Internet of musical things: Vision and challenges. IEEE Access, 6, 61994–62017 (2018)
15. Turchet, L and McPherson, A and Barthet, M: Co-design of a Smart Cajón. Journal of the Audio Engineering Society. 66, 220–230 (2018)
16. Berthaut, F and Marshall, M and Subramanian, S and Hachet, M: Rouages: Revealing the mechanisms of digital musical instruments to the audience. In Proc. NIME (2013)
17. Zappi, V and Mazzanti, D and Brogni, A and Caldwell, DG: Design and Evaluation of a Hybrid Reality Performance. In Proc. NIME, 355–360 (2011)
18. Mazzanti, D and Zappi, V and Caldwell, DG and Brogni, A: Augmented Stage for Participatory Performances. In Proc. NIME. 29–34 (2014)
19. Barthet, M and Fazekas, G and Sandler, M: Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. In Proc. CMMR. 492–507 (2012)
20. Malloch, J and Sinclair, S and Wanderley, MM: A network-based framework for collaborative development and performance of digital musical instruments. In Proc. CMMR. 401–425 (2007)
21. Mou, W and Gunes, H and Ioannis, P: Alone vs In-a-group: A Multi-modal Framework for Automatic Affect Recognition. ACM Transactions on Multimedia Computing, Communications and Applications (2019)
22. Juslin, PN and Lindstrm, E: Musical expression of emotions: Modelling listeners' judgements of composed and performed features. Music Analysis, 29(1/3), 334–364 (2010).
23. Braun, V and Clarke, V: Using thematic analysis in psychology. Qualitative Research in Psychology. 3(2) 77–101 (2006)

Designing Virtual Soundscapes for Alzheimer's Disease Care

Frédéric Voisin[1]

[1] Freelance
fred@fredvoisin.com

Abstract. Sound environment is a prime source of conscious and unconscious information which allows listeners to place themselves, to communicate, to feel, to remember. The author describes the process of designing a new audio interactive apparatus for Alzheimer's care, in the context of an active multidisciplinary research project led by the author in collaboration with a longterm care centre (EHPAD) in Burgundy (France), a geriatrician, a gerontologist, psychologists and caregivers. The apparatus, named *Madeleines Sonores* in reference to Proust's *madeleine*, have provided virtual soundscapes sounding for a year for 14 elderly people hosted in the dedicated Alzheimer's unit of the care centre, 24/7. Empiric aspects of sonic interactivity are discussed in relation to dementia and to the activity of caring. Scientific studies are initiated to evaluate the benefits of such a disposal in Alzheimer's disease therapy and in caring dementia.

Keywords: Sound design, Soundscapes, Sonic Interaction, Cognitive Rehabilitation, Alzheimer's Desease, Dementia, Caring, Quality of Life, Virtual Reality.

1 Introduction

Nowadays, Alzheimer's disease (AD) is the main cause of dementia. AD usually starts slowly and gradually worsens over time with brain damages and main effects on memory, cognition and behiavior : short-term and procedural memories are progressively affected until severely damaged, with severe spatio-temporal disorientation, when long-term memory becomes impaired later [1][2]. When different hypothesis may explain the causes of AD, there is no known validated pharmacologic therapy. Nevertheless, the research effort present promising results. Non-drug therapies are adapted to flatten or compensate the AD effects by stimulating cognitive and sensorimotor activities such as music practice and dance, which convene neural plasticity processes [3]. Playing and listening to music all together stimulate social forms of cognition, emotion and participate to re-entrainments of implicit and procedural memories affected by AD [4][5][6].

In the context of longterm care centres (EHPAD), these recent results encourage group musical activities led by music therapists in a variety of non-drug therapies for AD. Nevertheless in such a longterm care context, musical activity may be too rare and the benefits for each victim of AD are difficult to evaluate, particularly when the latter may present different forms of dementia.

As a geriatrician may observe in his own practice in such a context, too little attention is generally given to all sonic interactions which include not only music but various oral communications, noises, audio productions... and silences, all that actually defines the standard soundscapes of a living and social place in a hospital environment.

Moreover, such sonic interactions are able to recall facts from subjects' trivial (long term) memory, and reach the subcortical circuits spared by the disease, in particular those that are related to emotion as well as hearing.

In 2016, an experimentation by Dr Jeannin with AD victims showed that some peculiar sounds can *help recall long-term memories and emotions, when the listener's performance shows no relation with standard cognitive tests such as MMSE and NPI*. Dr Jeannin also demonstrates how long such an experience of auditory memory with simple sounds implies a knowledge of the biography of the listener : this biographical knowledge may facilitate the activity of caring by, for instance, sparking some *conversation* [7].

Therefore, some hospital's inner sound environment may not be well adapted to AD victims : during the day, loud radio or TV-shows play for a long time in individual or collective rooms and can make a continuous flow of informations that has mostly become incomprehensible in the studied cases of dementia. At night, wanderings are not rare in deep silence with no landmarks, no surprise, nowhere to go when the way back may be already lost in anonymous and closed halls, etc.

When sonic interactions may focus on the caring activity in a clinical perspective, a question may be: how would sound directly take part in AD caring ?

It is with this prospect that the Grégoire Direz Residence (EHPAD) in Mailly-le-Château has welcomed, since July 2017, a sound design research action in their unit specially equipped for AD victims. The hospital, its geriatric physician (Dr Pierre Jeannin), and its care team work together with a gerontologist (Prof. France Mourey[1]) on scientific aspects while I work on conceiving and carrying out the sound device on location: the « Madeleines sonores ».

To be pragmatic, we presume that from a functional and systematic point of view, not only music but every sound may involve various cognitive and emotional circuits [9][10]. An hypothesis would be that as soon as emotions and long-term memory still strong in later stages of AD (alexithymia appears late), not only music but any sound, if well chosen sounds, may help the caring activity [7] for AD : being *in situ* (i.e. on location), attention, social processes as well as emotion – itself able to move us – can be called upon as much through the music as from a variety of familiar or merely recognized sounds. Where, when and how sounds occur seem to be relevant questions to AD and dementia in a longterm health care centre.

Finally, one may consider that a continuity exists between "music" and "non-music" or "noise", which vary according to the human society and its history.

[1] Gerontologist, at Cognition, Action and Sensorimotor Plasticity Lab., UMR Inserm Unit 1093 - Université de Bourgogne, Dijon (France).

[2] I would like to thank Arnaud Bidotti, Master in Occupational and Organizational Psychology who led this study in March and April 2019 under the scientific supervision of Prof. Edith Salès-Wuillement, Laboratoire Psy-DREPI EA-7458 (Psychologie: Dynamiques Relationnelles et Processus Identitaires), Université de Bourgogne.

According to Luigi Russolo, Edgar Varèse, John Cage, Pierre Henri, Muray Schafer and others, a silent rest, a siren and other urban or natural soundscapes may be considered as "music" in certain conditions. Anthropologists and musicologists have shown a very large variety of music productions in relation to humans, animals, nature or surnatural beings.

2 Designing Soundscapes for AD victims

Pragmatically, we suggest designing virtual soundscapes adapted not only to different stages of dementia due to AD, i.e. to different mind states and beliefs, but also to some of the physiological effects of aging, such as a reduced visual field, hearing loss (or blocked ears), altered movement control, relative perceptions of time, etc. Starting from my own experience, from philosophical and ecological aspects of soudscapes from philosophical and theoretical frameworks devoted to soundscapes and mental health [8][9][10], and from rare references to some experiments on sound and soundscapes for AD victimes [11][12], I decided at first to chart the time and space of this very particular place and, then, to focus on anxiety phenomena that a sound environment may generate in certain conditions, particularly with elderly people.

After my first observations done for nights and days in the Alzheimer's Unit at the EHPAD under the supervision of Dr Jeannin and of the caring team, I decided at first to design the common areas of the Unit, where people spend place most of their time standing, moving, doing some activity or resting, with the care team.

The first sounds I had to produce had to obviously be congruent with the physical environment: big rooms, corridors, impersonal furniture, windows overlooking beautiful gardens and, in the distance, a village ... They also had to be congruent with the *immediate* perception and beliefs the residents may have, residents who are affected by (various forms of) dementia but who also bravely bear the various marks of now ancient times! To answer these multiple deteriorations of perception (vision, audition, proprioception), I deemed necessary to start with reinforcing what already existed to increase perception, thus “reality” in some ergonomical perspective. This is the context in which, afterwards, sounds were supposed – tactfully – to encourage each one’s imagination of former memories, collectively as well as individually, ordinary as well as extraordinary : not only with sound samples played out of their context, but sound effects.

A great variety of sound effects may partake in the staging of different sound realities and their modelling. With that prospect, I suggest understanding the notion of sound effect with an emphasis on a “logic of meaning”, between the cause and the event itself, “in a dimension that is altogether that of event and of situation”, through the “demonstration of a phenomenon that supports the [assumed] existence of the object” [13]. Therefore, the staging of this particular hospital space, that has become the canvas of a "film for the ear”, must be as much part of the scripting as of the content of the generated sound flows. By the way, the digital audioprocessing used in the actual disposal are not numerous: some peculiar synthesisers and peculier sound players with volume controls, fading, filters and a reverb. On the contrary, the logical

and stochastic effects are much richer and are the subject of the original computer coding (see below Section 4).

In the map above (see Figure 1), sounds produced from realistic sources from outside are supposed to be welcome in such a hospital's closed space : at its extreme borders, synthesized sounds of (virtual) bells give night and day to the listeners the (real) hours... or a landmark, as well as a short but regular game for cognitive stimulation (counting the strokes). In the patio, a granular audio synthesizer produces a continuous but varying waterfall sound, congruent with the beautiful but silent view, when the windows cannot be opened for security reasons.

This very first signals appear to be strong enough as landmarks to allow some algorithmic variations, depending on a mix between the actual and virtual place, on both the weather and the rhythm of the caring activity : when the sound quality of the waterfall or of the bells can change according to the outside temperature, or to the amount of humidity in air, their loudness intensity varies according to the care activity as well as to the cosmological events. Periods of days and seasons are also marked by a variety of geophonic sounds such as rain and wind which vary according to the actual weather... On the biophonic side, animal populations and their audio activity also change according the actual season, weather and ethological data. In the same way, the biophonic production is in relation with the precise location of the broadcasting of sound: birds that might come – whether in reality or virtually – to the waterfall in the patio are not the same ones as those that can be heard in the little living room on the North side, neither are they the same as those that can be heard in the big living on the South side.

The time-length of the animal sound sequences is approximately twenty minutes long, depending on weather conditions. Their intensity and density, therefore the duration of silences, depend on season and weather. Thus different tones continuously blend while they « naturally » diversify, as the algorithm gives the sound of birds in the rain, that would be more often crows in winter and tits in spring. When the sonorous broadcasting through loudspeakers is allowed for by the layout, audio spatialization effects may be produced, with moderation, to render movements coming from the wind, trees, birds, mammals...

Sounds that are distinctive to human activity, more variable are arbitrary, are broadcast in other (more infrequent) places of the common space. For example, in the area of the dining room sounds may suggest meal preparation. In this case, aiming at realism, and especially at respecting a multimodal congruence (here audition vs. taste), the choice of sound samles can, eventually, be made consistent with the menu, and the sound of a pressure cooker made become more frequent in winter than in spring: this implies, therefore, collaborating with cooks !

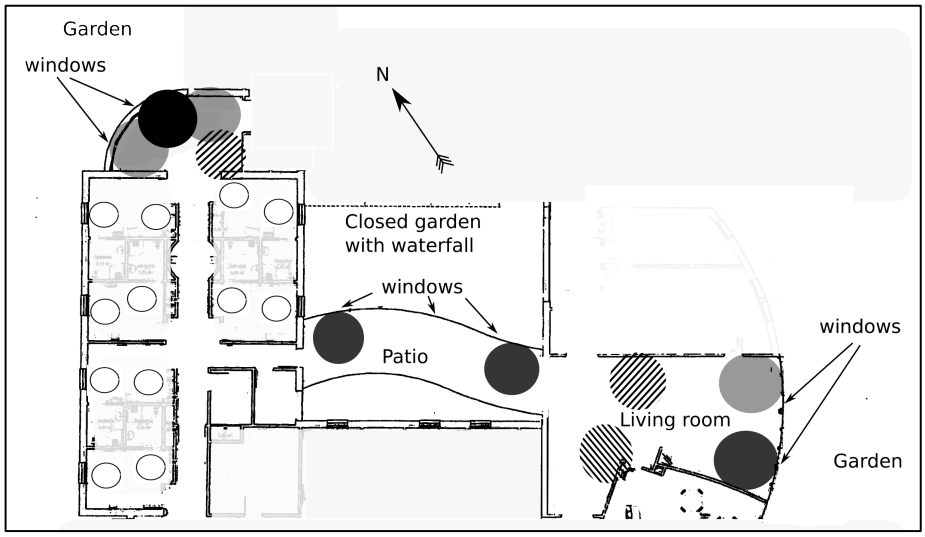


Figure 1. Layout of the loudspeakers in the common spaces of the Alzheimer's Unit: 1) the broadcasting points of the landmarks: bells and waterfalls (*black circles*), can also play geophonic and biophonic sequences ; 2) the broadcasting points of sounds for geophony or biophony and occasional human sonic activities from outside (*grey circles*) ; and 3) the broadcasting points of human audio activities, mostly from inside (*dashed circles*). The bedrooms have two broadcasting points each (*smaller empty circles*) : only six of them are shown on this map.

Extreme attention is given to the sound pressure intensity. I adapted the sound production with a rather unusual but conscious concern for the continuously exposed audience. Calibrated measures of acoustic pressure are underway.

In the common spaces, that are the noisiest, the acoustic pressure produced is unevenly distributed. Periods of silence (the acoustic pressure being of roughly 30 dBA) are frequent, except in the patio that overlooks the inner garden and its waterfall whose sound is continuous, though variable during long periods depending on time and weather conditions (minimum pressure of roughly 40 dBA). From the little living room on the north side, some sounds are broadcast loud enough to be heard as we arrive from the corridors that lead to it. In this more isolated place than the others, that leads to meditation or strolling, periods of silence alternate with active periods (with an acoustic pressure of 70 dBA and peaks at 80 dBA).

Until now, as is confirmed by the written notes taken by the care team, voiced complaints have never been about sound pressure, but rather about their eventual existence and pertinence. On the contrary, it even seems as if, with time, sounds tend to not be noticed consciously, at least by the health care workers. In agreement with the care team, the general sound pressure has even been reinforced with experimentation, in particular to compensate the eventual hearing deficiencies of the residents.

As recommended by music therapists [14], the general shape of a sound sequence, aside from the sound landmarks, for example the chirping of a bird, or horses passing,

etc. follows as much as is possible a U-shape for about 20', most often followed by a period of silence.

Insofar as we offer to encourage mnesic recalling and, in order to achieve that, to resort to attention cognitive processes, we suggest also resorting to a form of ABA' shape in which the sound shape A would tend to ressemble a call. It can be noticed, as a matter of fact, that the U-shape, that can also change into an inverted "J", conjures up a kind of returning (if only that of an increasing tempo), like the ABA' shape.

Each broadcasting point being independent, several sequences can partly overlap in various places. The number of simultaneous voices in the same location remains, however, voluntarily limited so as not to saturate the sound space, without, for that matter, preventing more or less fortuitous sounds meeting. Silences also allow, depending on what sounds are conjured up by the context, various effects that can concern sometimes a feeling of expectation (calls), at other times relief (response), mirroring the shape of a rhythm.

When the soundscapes of the common spaces take their inspiration from an as common reality, the broadcasting of soundscapes in the private spaces (bedrooms) is much more personalized. The scripting strategies there are somewhat different since the idea is to conjure up the residents' old memories while avoiding to cause confusional states that would take place in particular moments that the care team would help define. The choice of sounds and their staging here has to answer the congruence with a reality perceived in a state of dementia – which needs to be diagnosed in the best possible way – as well as the life story of the listener. We also have to admit that such diagnoses are not often enough made, with regards the evolution of the disease. Different protocols, based on the above-mentioned principles, are being considered.

Some simple and repetitive sounds, that conjure up a bygone past, may have a soothing effect: an initial observation, along with the care team, showed that the warm, slow and regular sound of a grandfather clock's pendulum, slowed down to an every-two-second impulse, seems to ease the sleeping phases and allow for a better sleep. Such observations are still, at this point of experimentation, to be confirmed experimentally: the rooms are in the process of being empirically and progressively fitted with sound. We hope that, in the long run, such an experimentation may enable to elaborate a precise questionnaire, based – for example – on a guided listening of sounds which are defined beforehand, that better allows to characterize each resident's sensitivity to different sounds.

This personnalized aspect of conception, the most delicate, is in the course of being experimented and will progressively be deployed.

3 Soundscapes for caregivers

Obviously, some sounds may be difficult to play with. In some conditions, with dementia, when the perception of some sounds imply, as a reflex, the viewing of their origin, their separation from their real or imaginary cause may create confusional states: to avoid such a case, the help of a caregiver is useful to explain the origin of such a sound, to calm the person down and to finally have — as expected — some *conversation*.

Conceiving a sound environment and its dynamics, necessariy involves the health care team. Sound must not interfere with care, it must help and support it. We can therefore consider that sound may help the patient in directly adressing him or her while addressing the care team as well. With their assent and their complicity, oral conversations may be engaged in and sounds adopted.

For example, in the very first days following the beginning of the setup (in July 2018), nursing auxiliaries asked me for the new bells in the commons spaces never to leave: the opportunity and ergonomics of such a multi-functional sequence of a simple sound, repeated twice an hour (at 2' interval), 24 hours a day allows to mark a territory, to recall an hour common to all, to be the opportunity for a social exchange, and to become a training of different short-term memories, indeed to become part of spontaneous cognitive assessment, almost at any time. And slight periodic variations of the sound synthesis of bells can be sought for, for who might want it if only per chance. More or less shared knowledge, whether popular, personal or specialized, may be called upon on the subject of some animal sound production. Producing some insect sounds, or mammal ones, for example, has been the subject of discussions with the care team as regards their realism, the cridibility, or the emotional states they are likely to cause.

Soundification of virtual human activities may also take part in the organisation of care: for example, the sound accompanying meals may be realised in such a way that it more or less consciously prepares the residents to eat. This anticipation may be all the more appreciated as the residents themselves are incapable, because of their dementia, of knowing that it is *in fact* really time to eat... Such an effect is most certainly welcome when it contributes to the support of demented people in the absence of any other clue that would be obvious and familiar to them, in a hospital context.

The developpement of a closer feedback and control interface from the care team is contemplated. Conceiving such an interface would should directly integrate existing computer interfaces already devoted to caring. With that as a goal, research in psychosociology on the links between sonic environment and care has already began under the supervision of Prof. France Mourey. A first study of the effects of “Madeleines sonores” (the virtual soundscapes, the plan of action) on the relationship between caregivers and care receivers is based on collecting information during interviews with the various professionals linked with this project. The thematic analysis [15] of the interviews that were carried out have allowed to establish one (or more) questionnaire(s) that may characterize the representations and attitudes that the care team takes up faced with the sound plan of action[2]. We are considering that the use of such a questionnaire may allow to know the attitudes that come into play while applying the sound plan of action, this being to clarify the contribution in relation with the resident. Moreover, knowing how the care team sees this sound plan of action also partakes in evaluating its effects, in terms of contribution or benefits, with the residents.

[2] I would like to thank Arnaud Bidotti, Master in Occupational and Organizational Psychology who led this study in March and April 2019 under the scientific supervision of Prof. Edith Salès-Wuillement, Laboratoire Psy-DREPI EA-7458 (Psychologie: Dynamiques Relationnelles et Processus Identitaires), Université de Bourgogne.

4 A Digital Architecture for Virtual Soundscapes

The soundscapes are systematically generated and mixed in real-time using a distributed architecture with a principal generator (see the « Soundscape Generator », in Figure 2 below) that can be described as a server hosting the sound database and a set of routines periodically executed by its operating system scheduler[3]. These routines, written in Lisp language and Unix-like shell scripts, compute different kinds of mapping, empirically developped, relevant to the environmental data as modelised, to its inner virtual population and to individual data (biographies). They finally control 24 on-board computers[4] fixed in the ceiling of the common and private spaces and hosting their own digital signal processing (DSP) and loudspeakers (the « Players », in Figure 2).

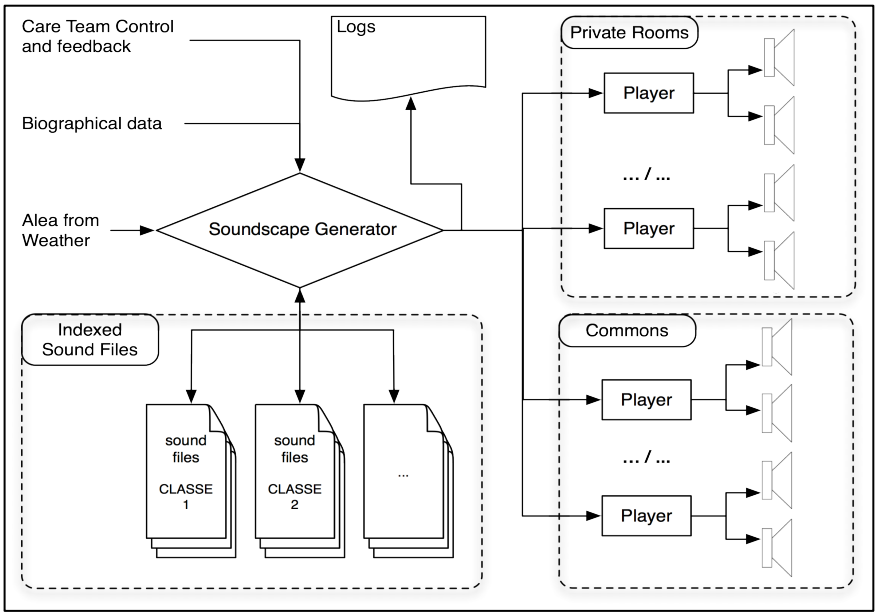


Figure 2. General principle of the *Madeleines Sonores*

All the soundscapes are generated as a remix of short sound samples (lasting from a few seconds to a few dozens of seconds), or using sound synthesizers models adapted from my own experience [16] and from [17]. The sound samples, specially recorded or adapted from open sound collections to represent various places, activities, animals and human societies from the 40's, 50's, are retrieved from a large

[3] Cf. cron programs in Unix-like operating systems (see: https://en.wikipedia.org/wiki/Cron).

[4] All computers operating systems are GNU-Linux based. The on-board ARM based computers are Raspberry Pi3-B+ with audio boards IQaudio Pi-DigiAMP+ and in-ceiling 6.5" loudspeakers QI 65C/St by QAcoustics (70 Hz to 16 kHz ± 3 dB, up to 90 dBA max pressure level). The DSP is built using Pure Data open-source software [18]. The communications are based on TCP/IP protocols: secure-shell (SSH) for asynchronous sound transferts, UDP for real-time controls.

database with thousands of items indexed by myself. Periods of silence between sound samples or synthesizers are more or less short and always irregular. They define the density of virtual events that occur depending on each resident's sensitivity and the general context: according to daytime periods, the year, the care provided, etc.

In the individual rooms, the soundscapes are rendered using two equivalent monophonic loudspeakers placed across the room, one close to the window and the other facing the bed. The activation of the soundscapes entirely depends on the explicit approval of whom concerned: the individual, himself, herself or their family when he or she cannot agree for clinical reasons and, obviously, by the caring team (doctors, psychologists, nurses).

All hi-level instructions, play-lists, sound filenames, pressure levels, etc are systematically locally logged. The overall disposal can be remotely controled and monitored thru a standard secured connection (SSH) on internet, using any Unix-like terminal. Confidentiality is ensured in particular by the entire anonymity of the digitized data.

5 Discussion

The work presented here is only the beginning of a long conceiving process in an emerging activity within the field of sound design, mental health and care. Its actual state is an action-research led in the fields of sound art, human and health science, and technology.

Our first observations are encouraging: the behaviour of part of the residents suggest that not only the residents appreciate the variety of the new sound environment but they tend to place themselves close to the audio sources to recollect or to meditate, whether consciously or not (see Fig. 3 below). The questionnaire we developed should make it possible to make care givers aware of the effects of sound environment on AD victims.

After such a first research that was necessarily empirical, considering the innovative characteristic of the project, applications respecting a truly experimental protocol have to be considered.

Optimizing the algorithms that was developped for that purpose still has to be done, as the deployment of the plan of action is being ratified. Particular attention should be focused onto indexing the sound samples insofar as they should enable, not only to identify the origin, the nature and the quality of sounds, but also, beyond their psycho-accoustic properties, the emotions and reactions they are liable to cause, as well as the broadcasting methods (context). In fact, we are contemplating the fact that, ultimately, the care team may report the reactions they provoke thanks to an interface that remains to be created. Such an interface would then enable to directly partake in the experimental phase of the plan.

The whole plan of action is for the moment operating as a multimedia work of art (sound and computer) that falls within the altogether artistic, scientific and clinical perspective that federates different social and economic players in a humanistic dynamic of research and innovation.

Figure 3. A resident is meditating at the closest place to the audio source mounted at the ceiling.

Thanks

I am particularly thankful to Dr Pierre Jeannin for his benevolence, his logistic and voluntary support, and of course to the charity Castel-Mailletaise that welcomed this project, and the care team that is closely involved in it. My thanks also go towards France Mourey for her support. And I am grateful to Claire Webster for emending and translating parts of this article.

References

1. Eustache, F. : Langage, vieillissement et démences. In F. Eustache & B. Lechevalier (Eds.) *Langage et Aphasie, Séminaire Jean-Louis Signoret,* 205–228. Bruxelles : De Boeck Université (1993)
2. Dubois, B., Feldman, H. H., Jacova, C., Cummings, J. L., DeKosky, S. T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N. C., Galasko, D., Gauthier, S., Hampel, H., Jicha, G. A., Meguro, K., O'Brien, J., Pasquier, F., Rober, P., Rossor, M., Salloway, S., Sarazin, M., de Souza, L. C., Stern, Y., Visser, P. J., & Scheltens, P. : Revising the definition of Alzheimer's disease : a new lexicon. *Lancet neurology 9*, 118–127 (2010)
3. Chancellor, B., Duncan, A., Chatterjee, A. : Art Therapy for Alzheimer's Disease and Other Dementias. *Journal of Alzheimer's Disease 39*, 1–11 (2014)
4. Guétin, S., Portet, F., Picot, M.C., Pommié, C., Messaoudi, M., Djabelkir, L., Olsen, A.L., Cano, M.M., Lecourt, E., Touchon, J. : Effect of Music Therapy on Anxiety and Depression in Patients with Alzheimer's Type Dementia: Randomised, Controlled Study. *Dementia and Geriatric Cognitive Disorders 28*, pp. 36–46 (2009)
5. Moussard, A., Bigand, E., Clément, S., Samson, S. : Préservation des apprentissages implicites en musique dans le vieillissement normal et la maladie d'Alzheimer. *Revue de neuropsychologie 18*, 127–152 (2008)
6. Mofredj, A., Alaya, S., Tassaioust, K., Bahloul, H., Mrabet, A. :2016. Music therapy, a review of the potential therapeutic benefits for the critically ill. *Journal of Critical Care 35*, pp. 195–199 (2016)
7. Jeannin, P. : Projet Hippocampe : Les madeleines sonores. Essay, http://fredvoisin.com/IMG/pdf/pierre_jeannin-projet_hippocampe.pdf (2017)
8. Schafer, R.M. : *The Tuning of the World*, Knopf, New-York (1977)
9. Westerkamp, H. : Linking Soundscape Composition and Acoustic Ecology. *Organised Sound*, 7/1 (2002).
10. Andringa, T., Lanser, J. : How Pleasant Sounds Promote and Annoying Sounds Impede Health: A Cognitive Approach. *International Journal of Environmental Research and Public Health 10*, 1439–1461 (2013)
11. Nagahata, K., Fukushima, T., Ishibashi, N., Takahashi, Y., Moriyama, M. : A soundscape study : What kinds of sounds can elderly people affected by dementia recollect? *Noise and Health 6*, 63 (2004)
12. Burgio, L., Scilley, K., Hardin, J. M., Hsu, C., & Yancey, J. : Environmental «white noise»: an intervention for verbally agitated nursing home residents. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences,* 51(6), P364-373 (1996)
13. Augoyard, J.F., Torgue, H. : *Sonic experience: a guide to everyday sounds*. McGill-Queen's Press-MQUP (2014)
14. Guétin, S., Portet, F., Picot, M.C., Pommié, C., Messaoudi, M., Djabelkir, L., Olsen, A.L., Cano, M.M., Lecourt, E., Touchon, J., 2009. Effect of Music Therapy on Anxiety and Depression in Patients with Alzheimer's Type Dementia: Randomised, Controlled Study. *Dementia and Geriatric Cognitive Disorders 28*, 36–46. https://doi.org/10.1159/000229024
15. Negura L. : Content analysis in the study of social representations (L'analyse de contenu dans l'étude des représentations sociales), *SociologieS* [Online], Theory and research (2006), connection on 2019 May 9th, http://journals.openedition.org/sociologies/993
16. Author's website, http://www.fredvoisin.com/
17. Farnell A. : *Designing Sound*. The MIT Press. Cambridge, Mass. (2010)
18. Puckette M. : Pure Data, *Proceedings of the International Computer Music Conference (Hong Kong, 1996)*, ICMA, 269-272 (1996).

ARLooper: a Mobile AR Application for Collaborative Sound Recording and Performance

Sihwa Park

Media Arts and Technology
University of California, Santa Barbara
sihwapark@mat.ucsb.edu

Abstract. This paper introduces ARLooper, an AR-based iOS application for multi-user sound recording and performance, that aims to explore the possibility of actively using mobile AR technology in creating novel musical interfaces and collaborative audiovisual experience. ARLooper allows the user to record sound through microphones in mobile devices and, at the same time, visualizes and places recorded sounds as 3D waveforms in an AR space. The user can play, modify, and loop the recorded sounds with several audio filters attached to each sound. Since ARLooper generates the world map information through iOS ARkit's tracking technique called visual-inertial odometry which tracks the real world and a correspondence between real and AR spaces, it enables multiple users to connect to the same AR space by sharing and synchronize the world map data. In this shared AR space, the user can see each other's 3D waveforms and activities, such as selection and manipulation of them, as a result, having a potential of collaborative AR performance.

Keywords: Augmented reality · AR · Mobile music · Mobile instrument · Mobile performance · Collaboration

1 Introduction

Mobile computing and technology have affected the field of computer music in terms of the exploration of musical interfaces and performance paradigms. Before smartphones with a touchscreen emerge, the relevant beginning of this history can be referred to a body of pioneering work that experimented the active use of mobile phones, which were just ordinary consumer electronic devices, in creating interfaces for musical performances that allow audiences to participate in a part of performance or making new forms of musical instruments that expand performers' musical expression. Along with these experiments and applications of mobile computing in music technology, a concept of mobile music was also suggested by Gaye et al. [1], emphasizing the mobility and collaborative aspects of mobile music. Since the emergence of smart devices such as smartphones and tablets, there has been various research that explores possibilities of using consumer electronic devices as New Interfaces for Musical Expression (NIME) by utilizing diverse sensors embedded in these devices, such as GPS, cameras,

gyroscopes, accelerometers, touchscreens, wireless sensors, etc., expanding the field of mobile music technology. Many of examples are well documented by Essl at el. [25].

Mobile augmented reality (AR) technology has also evolved in this context of mobile computing [6]. Especially, advance in tracking physical objects and environments without any prior information and fusing virtual content with the real world without 2D fiducial markers or images has been significantly contributing to mobile AR research since seminal works of researchers including Davison [5], Klein and Murray [4] were introduced. Agilely adopting this technology, major companies such as Google and Apple have been leading this domain by releasing and updating development frameworks, e.g., ARCore and ARKit, for mobile AR applications that can be running on commonly accessible devices. This markerless mobile AR, however, has been yet not enough explored in the fields of mobile music technology and NIME. It has a potential to devise a novel and collaborative interface for musical creativity in an AR space and create unique musical experience for both performers and audiences.

As an initial study on this promising exploration, this paper introduces ARLooper, an AR-based mobile interface for multi-user sound recording and performance. ARLooper allows the user to record sound through microphones in mobile devices and, at the same time, visualizes and places recorded sounds as 3D waveforms in an AR space. The user can play, modify, and loop these recorded sounds with several audio filters attached to each sound. ARLooper employs ARKit's tracking technique, which is called visual-inertial odometry to track the real world and a correspondence between real and AR spaces by combining information from camera sensor data with the device's motion sensor data, so that the multiple users of ARLooper can connect to the same AR space in which recorded sounds are shared and can see each other's activities, such as selection and manipulation of sound waveforms.

2 Literature Review

Considering that ARLooper tries to embrace the possibility of the mobile AR in mobile music technology, it is reasonable to review the context of prior research for mobile phone-based music making and collaborative performance.

In addition to Levin's Dialtones (A Telesymphony) [7] considered as one of the pioneering works that exploits the mobile phones of the audience as a part of musical performances, the early attempts to re-purpose not only a variety of built-in sensors of consumer mobile electrics but also the mobility of the devices for NIME have been actively made and well documented by several researchers [1][2].

On top of the focus of using mobile devices in making musical interfaces, paradigms for collaborative mobile music-making and performance also has been explored and suggested. Tanaka [8] created an early system with PDAs of which performers collaborate in controlling audio streams over mobile wireless networks, thereby creating a piece of music together in real time. Performers of

CaMus2 [9], which used the built-in camera of mobile phones to track maker sheets as visual references for musical interaction, shared their interaction parameters through Bluetooth communication to control and generate sound by transforming sensor data into MIDI messages. While these systems created sound in remote computes, the Stanford Mobile Phone Orchestra (MoPho) [10] took the advantage of iPhone's capability in synthesizing sound in local mobile devices by running the custom software MoMu Toolkit, introducing a new paradigm for mobile phone performances from their own perspective. Alessandro [11] and Lee et al. [12] also demonstrated a possibility of collaborative music-making with mobile devices by utilizing networking functionality. Salazar et al. [3] documented various performances using Auraglyph, a sketch gesture-based iPad music application, as the extended paradigm of mobile music performance.

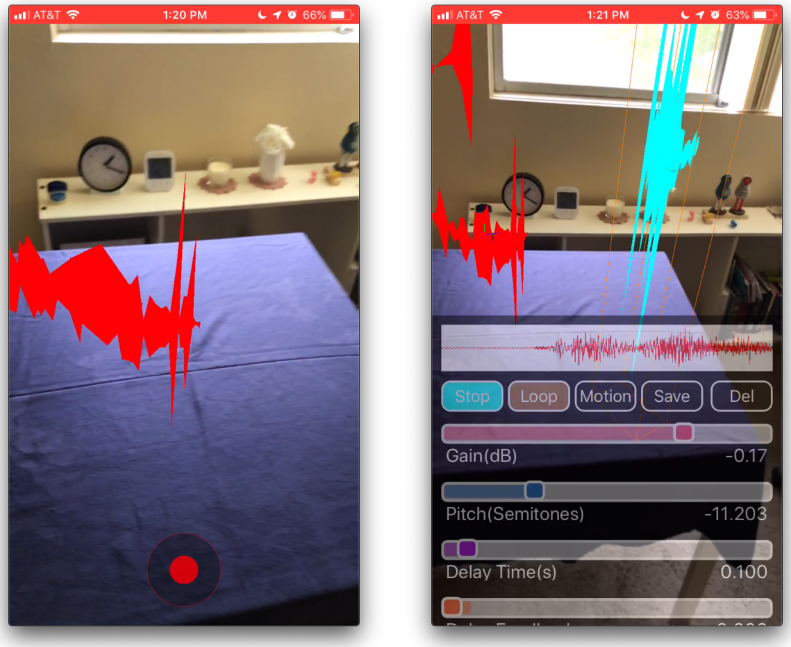
It is also noteworthy to recognize how the AR technology has evolved to support the environments for Computer Supported Cooperative Work (CSCW). Since the first collaborative AR projects, such as the Studiersube [13] and Shared Space [14], emerged in the mid-nineties, there have been a lot of research and examples of collaborative AR applications that support remote or co-located user contexts as it can be seen in Zhou et al.'s review [15]. One of interesting advances is that fundamental environments for collaborative AR research have changed from surface projection-based systems, e.g., Rekimoto and Saitoh's Augumented Surfaces [16], to handheld or wearable devices with see-through displays, e.g., Emmie [17] and ARTHUR [18], thereby enabling 3D virtual information to be overlaid anywhere in the real world in ways that would be otherwise impossible.

Compared to these abundant AR research for CSCW, a few AR research has been conducted for Computer Mediated Musical Collaboration (CMMC) [19]. Poupyrev et al. [20] presented Augmented Groove, an AR musical interface using head-mounted displays and 2D fiducial markers, of which the user play music together by controlling physical cards where 3D virtual controllers and images are attached to. Similar to Augmented Groove, Berry's The Music Table [21] used marker cards as musical sources in a compositional process in which the user arranges the cards on a table. While these systems need to share the same physical markers on a single table, YARMI [22], a multi-user, networked musical instrument, operates on multiple tabletops based on client-server architecture, presenting a shared AR space among performers as the concept of synchronized ensemble. In terms of augmenting physical objects with virtual information, the reacTable [23] also could be considered as one of AR musical systems that suggested networked multiple tabletop interfaces for collaboration, although it only overlays 2D graphics below the objects placed on tabletop screens. As a descendant of the reacTable, Clouth [24] created its AR version that works on mobile devices, attempting to use AR as a control mode for the reacTable but without a feature for collaboration.

3 Design

3.1 ARKit and AudioKit

ARLooper is an iOS application developed with ARKit[1], Apple's AR development toolkit for iOS devices, and AudioKit[2], a sound synthesis, processing, and analysis framekwork for the operating systems running on Apple's products, such as iOS, macOS and tvOS.



(a) Recordingn UI (b) Control UI

Fig. 1: ARLooper

3.2 AR-based Sound Recording and Visualization

ARLooper allows users to record sound incoming through the microphone of a device in real time by presenting a recording GUI button as shown in Fig. 1a. When the user presses the button, it immediately records a sound and at the same time, visualizes the currently recording sound as a 3D waveform in an

[1] https://developer.apple.com/arkit/
[2] https://audiokit.io/

AR space in a way that the height of 3D waveforms represents amplitude s of recorded sound samples. While the recording is in progress, the x and z positions of the 3D waveform are determined by the position of the device which the user holds and moves around in a space. This interaction has a metaphor of 3D drawing but with real-time audio data. As a result, the user can perform 3D drawing interaction in an AR space by easily understanding this real-time audio recording and 3D visualization. This recording process continues until the button is pressed again and is possible to repeat upon the user's intent, enabling the recording of multiple sound sources to be used later.

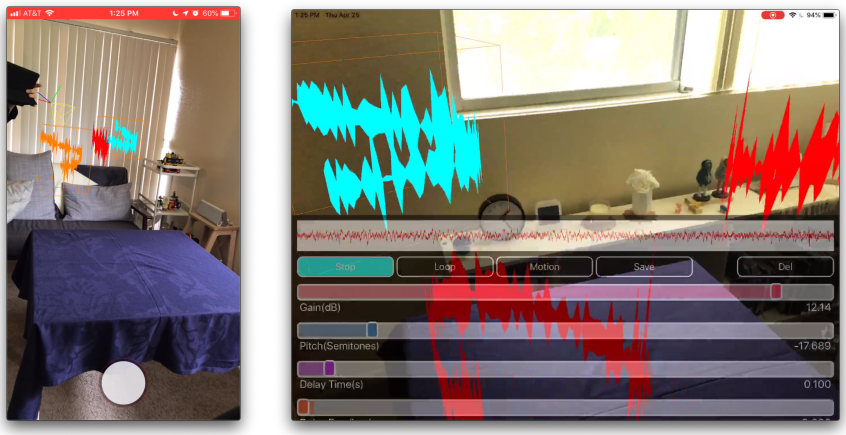
3.3 Sound Control

For audio recording, processing, and playback, ARLooper uses AudioKit's various sound filters and effects, such as a pitch shifter, a delay effect, a low pass filter, a reverb effect, a tremolo effect, and a gain control, which are attached to each recorded sound. When the user taps a 3D waveform, it presents a GUI to control the parameters of these attached filters and effects in addition to a conventional 2D waveform view of the recorded sound and buttons for playback, looping, and deletion (See Fig. 1b). The GUI disappears when the user re-taps the 3D waveform or taps a space where no AR waveform object exists. If a waveform is selected, its bounding box is shown in orange color to indicate that the waveform is selected. To distinguish the control status of a waveform, three colors are uses: red, orange, and cyan. Red is for ready after recording. Orange means it is selected, whereas cyan color represents it is playing.

On top of the GUI-based sound manipulation, ARLooper also gives a gesture-based interaction method on a touchscreen thereby supporting direct audiovisual manipulation. It has two types of gestures: pinch and rotation. When the user pinches a selected 3D waveform inward or outward, the size of the waveform decreases or increases, respectively. Along with the visual size change, its sound gain is simultaneously modulated in the same manner. In case of rotation, the pitch of the sound shifts according to the rotation of the waveform. As a result, with the gesture interaction, the user can more intuitively control the gain and pitch of a sound, not through GUI sliders.

3.4 Collaborative AR Mode

ARLooper utilizes ARKit's tracking technique, visual-inertial odometry that fuses information from camera sensor data with the iOS device's motion sensor data. It recognizes feature points in a visible scene captured from the camera of the device and as a result, infers the device's position and motion in a space. Based on this process, ARKit generates a world-mapping data called ARWorldMap that can be used as a shared frame of reference for a multi-user AR experience. To create a shared AR space among users, it uses the iOS Multipeer Connectivity framework that supports the discovery and data communication among nearby devices; One user starts as a host with ARLooper, finds a reference plane, and places a virtual plane with additional adjustment; Other users



(a) iPhone User's Screen (b) iPad User's Screen

Fig. 2: Multi-user Mode

then join a shared AR space as guests by requesting and obtaining an ARWorldMap instance from the host user. Here the guests have to find a similar view of the frame of reference in the space to re-localize to the received AR world map.

Once the AR session for multiple users is established, it is possible to share virtual content or information about user interaction. When a user creates a 3D waveform, its visual data is shared with others in real time so that everyone in the same AR space can see each other's recording activity (See Fig. 2). Also, information about the user's manipulation such as playback, parameter control, and gesture interaction is synchronized among all participants in the session. One important aspect in this shared sound AR space is a protocol for isolating the manipulation of a waveform according to who is a dominant user; ARLooper allows only one user to play a waveform at the same time by preventing others from having access to operate the waveform.

4 Discussion and Future Work

As the initial prototype of a long-term project, ARLooper shows the possibility of exploiting the recent mobile AR technology in the mobile music field and creating a novel and collaborative audiovisual experience in an AR space. Although the single user mode of ARLooper has been tested and used in a couple of actual performances by participating as a member of the UCSB Center for Research in Electronic Art Technology (CREATE) Ensemble[3], the multi-user mode still

[3] http://www.create.ucsb.edu/

needs to be more stabilized in sharing real-time visual and sound data. Considering that AR-based musical interfaces and performances are not common, this research should consider relevant use scenarios and a performance paradigm. Especially, since this AR technology is subject to a lighting condition in tracking objects from captured images, finding the best stage setup for performances will be one of the essential aspects to be deliberated. For a better user experience, it is also necessary to improve user interaction in manipulating AR waveforms; for example, a feature for loading and placing pre-recorded audio files in a AR space, an additional UI for selecting the playback range of a recorded sound in a 2D waveform view, and multi-touch based direct parameter control over AR waveforms. Reflecting that AR waveforms are positioned in a physical space, it would be interesting to spatialize recorded sounds in playback according to the direction and distance between AR waveforms and users. Also, it would be worth attempting to find an aesthetically better way to visualize recorded sound in 3D. Finally, it is imperative to conduct quantitative and qualitative user tests of ARLooper which will give invaluable insights to improve its usability.

References

1. Gaye, L., Holmquist, L.E., Behrendt, F. and Tanaka, A.: Mobile Music Technology: Report on an Emerging Community. In: the International Conference on New Interfaces for Musical Expression, pp. 22–25. Paris, France (2006)
2. Essl, G. and Rohs, M.: Interactivity for Mobile Music-making. Organised Sound, vol. 14, no. 2, pp. 197–207 (2009)
3. Salazar, S., Pipepenbrink, A. and Reid, S.: Developing a Performance Practice for Mobile Music Technology. In: the International Conference on New Interfaces for Musical Expression, pp. 59–64. Virginia Tech, Blacksburg, Virginia, USA (2018)
4. Klein, G. and Murray, D.: Parallel Tracking and Mapping for Small AR Workspaces. In: the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 1–10. IEEE Computer Society (2007)
5. Davison, A. J.: Real-time Simultaneous Localisation and Mapping with a Single Camera. In: Iccv, vol. 3, pp. 1403–1410 (2003)
6. Arth, C., Grasset, R., Gruber, L., Langlotz, T., Mulloni, A. and Wagner, D.: The History of Mobile Augmented Reality. arXiv preprint arXiv:1505.01319 (2015)
7. Levin, G.: Dialtones (A Telesymphony). (2001) (Accessed on 05.03.2019) `http://www.flong.com/projects/telesymphony/`
8. Tanaka, A.: Mobile Music Making. In: the International Conference on New Interfaces for Musical Expression, pp. 154–156. Hamamatsu, Japan (2004)
9. Rohs, M. and Essl, G.: CaMus 2: Collaborative Music Performance with Mobile Camera Phones. In: the International Conference on New Interfaces for Musical Expression, pp. 190–195 (2007)
10. Oh, J., Herrera, J., Bryan, N.J., Dahl, L. and Wang, G.: Evolving The Mobile Phone Orchestra. In: the International Conference on New Interfaces for Musical Expression, pp. 82–87. Sydney, Australia (2010)
11. d'Alessandro, N., Pon, A., Wang, J., Eagle, D., Sharlin, E. and Fels, S.: A Digital Mobile Choir: Joining Two Interfaces towards Composing and Performing Collaborative Mobile Music. In: the International Conference on New Interfaces for Musical Expression. University of Michigan, Ann Arbor, Michigan, USA (2012)

12. Lee, S.W., Srinivasamurthy, A., Tronel, G., Shen, W. and Freeman, J.: Tok! : A Collaborative Acoustic Instrument using Mobile Phones. In: the International Conference on New Interfaces for Musical Expression. University of Michigan, Ann Arbor, Michigan, USA (2012)
13. Szalavári, Z., Schmalstieg, D., Fuhrmann, A. and Gervautz, M.: "Studierstube": An Environment for Collaboration in Augmented Reality. Virtual Reality, vol. 3, no. 1, pp.37-48. Springer (1998)
14. Billinghurst, M., Weghorst, S. and Furness III, T.: Shared Space: An Augmented Reality Interface for Computer Supported Collaborative Work. In: Proc CVE, vol. 96. (1996)
15. Zhou, F., Duh, H. B. L. and Billinghurst, M.: Trends in Augmented Reality Tracking, Interaction and Display: A Review of Ten Years of ISMAR. In: the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 193–202. IEEE Computer Society (2008)
16. Rekimoto, J. and Saitoh, M.: Augmented Surfaces: A Spatially Continuous Work Space for Hybrid Computing Environments. In: the SIGCHI Conference on Human Factors in Computing Systems, pp. 378–385. ACM (1999)
17. Butz, A., Hollerer, T., Feiner, S., MacIntyre, B. and Beshers, C.: Enveloping Users and Computers in a Collaborative 3D Augmented Reality. In: the 2nd IEEE and ACM International Workshop on Augmented Reality, pp. 35–44. IEEE (1999)
18. Broll, W., Lindt, I., Ohlenburg, J., Wittkämper, M., Yuan, C., Novotny, T., Mottram, C. and Strothmann, A.: Arthur: A Collaborative Augmented Environment for Architectural Design and Urban Planning. Journal of Virtual Reality and Broadcasting, vol. 1, no. 1 (2004)
19. Arango, J.J. and Giraldo, D.M.: The Smartphone Ensemble. Exploring Mobile Computer Mediation in Collaborative Musical Performance. In: the New Interfaces for Musical Expression Conference, pp. 61–64. Brisbane, Australia (2016)
20. Poupyrev, I., Berry, R., Kurumisawa, J., Nakao, K., Billinghurst, M., Airola, C., Kato, H., Yonezawa, T. and Baldwin, L.: Augmented groove: Collaborative Jamming in Augmented Reality. In: ACM SIGGRAPH 2000 Conference Abstracts and Applications, vol. 17, no. 7, p. 77 (2000)
21. Berry, R., Makino, M., Hikawa, N. and Suzuki, M.: The Augmented Composer Project: The Music Table. In: the 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 338–339. IEEE (2003)
22. Laurenzo, T., Rodríguez, E. and Castro, J.F.: YARMI: An Augmented Reality Musical Instrument. In: he New Interfaces for Musical Expression Conference, pp. 268–269. Pittsburgh, PA, USA (2009)
23. Kaltenbrunner, M., Jorda, S., Geiger, G. and Alonso, M.: The Reactable*: A Collaborative Musical Instrument. In: the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 406–411. IEEE (2006)
24. Clouth, R.: Mobile Augmented Reality as a Control Mode for Real-time Music Systems. Universitat Pompeu Fabra, Barcelona (2013)
25. Essl, G. and Lee, S.W.: September. Mobile Devices as Musical Instruments-state of the Art and Future Prospects. In: International Symposium on Computer Music Multidisciplinary Research, pp. 525–539. Springer, Cham. (2017)

Singing in Virtual Reality with the Danish National children's choir

Stefania Serafin, Ali Adjorlu, Lars Andersen and Nicklas Andersen

Multisensory Experience Lab,
Aalborg University Copenhagen
sts@create.aau.dk

Abstract. In this paper we present a Virtual Reality (VR) system that allows a user to sing together with the Danish National Children choir. The system was co-designed together with psychologists, in order to be adopted to prevent and cope with social anxiety. We present the different elements of the system, as well as a preliminary evaluation which shows the potential of the system as a tool to help coping with social anxiety.

1 Introduction

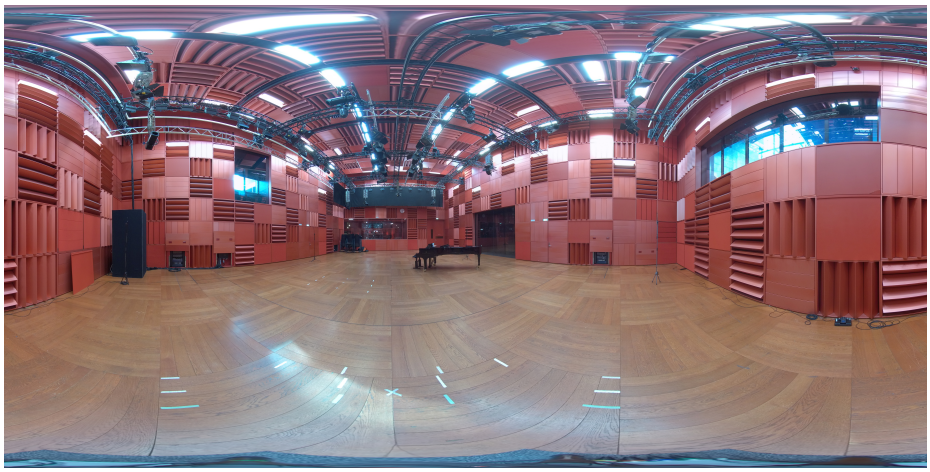


Fig. 1. A panoramic view of the studio at Danish Radio in Copenhagen, Denmark, where the recordings took place.

Social anxiety concerns the fear and subsequent avoidance of social situations. In a person suffering from anxiety, this negatively affects the participation in social situations (e.g., attending school), the forming of social relationships, and overall reduces quality of life. If the anxiety persists through adulthood, a person

is less likely to become a contributing member of society by finishing an education and getting a job. As a consequence, people suffering from anxiety tend to isolate themselves and live a life of lost opportunities.

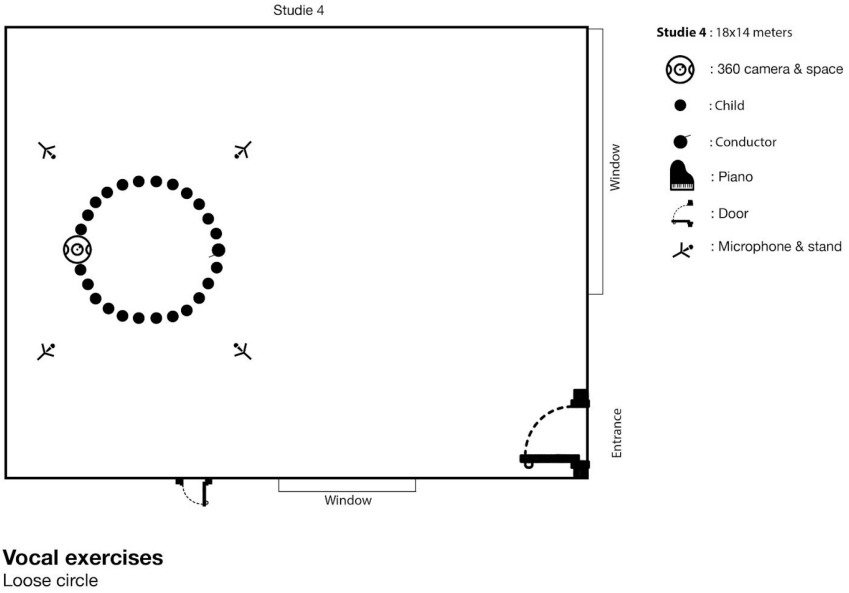


Fig. 2. The placement of camera, microphone and children during the vocal exercises: children placed in a loose circle.

A commonly used method for treating phobia and anxiety is exposure therapy, where the patient is exposed to the object that induces the anxiety. Exposure therapy has been shown to be effective with anxiety disorders, and was first used for social phobia in the mid-1980s. This is due to the fact that exposure to realistic social situations is very difficult to conduct and organise [2].

Lyneham and colleagues have developed a cognitive behavioral therapy tool for children called Cool Kids [9]. Cool Kids Program is a scientifically based and effective anxiety treatment program. The program was developed in Australia and translated into Danish at the Anxiety Clinic for Children and Youth at Aarhus University. In a Cool Kids group course, children and parents are introduced to concrete methods and strategies that can make it easier to overcome and deal with anxiety in everyday life. Cool Kids Program is based on principles of cognitive behavioral therapy and can be implemented either individually or in groups with other children and their parents. The program has been developed throughout more than 15 years of research and has some evidences that a fraction of the children completing the program are diagnosis-free or markedly improved.

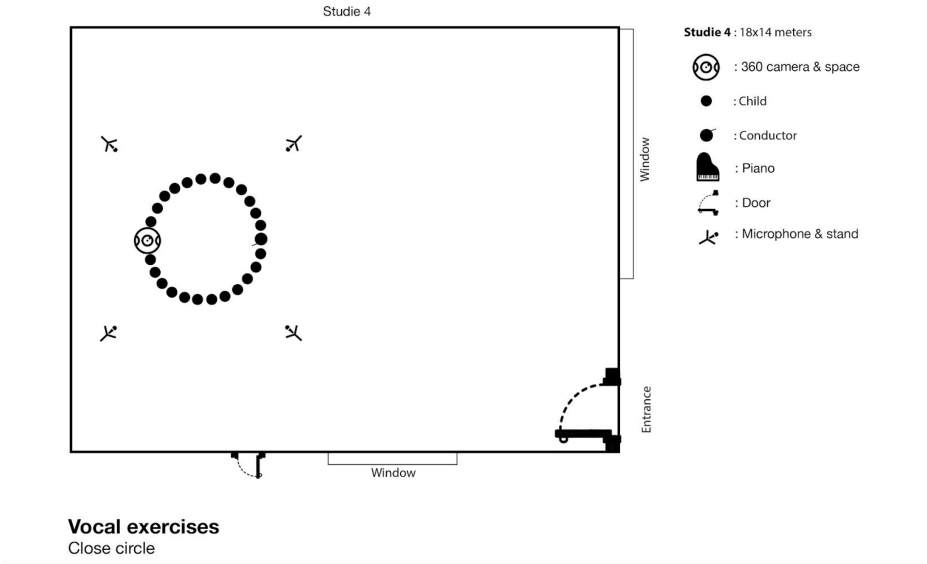


Fig. 3. The placement of camera, microphone and children during the vocal exercises: children placed in a closed circle.

Virtual Reality (VR) has become a popular platform used in interventions, and research has delivered successful results in areas including social and cognitive training exercises combined with exposure therapy [1, 7, 11, 12]. While VR headsets are not in most homes currently, the development is going in that direction due to lower cost and wider availability.

The use of VR in the treatment of anxiety is grounded in emotional processing; fear memories can be constructed as structures that contain information regarding stimuli, responses and meaning [5]. VR is presumed to activate the fear structure by immersing the individual in the feared situation and to modify the fear structure through the processes of habituation and extinction.

Since the use of VR in social and cognitive training is a relatively new area of exploration, to our knowledge there are no applications to music or sound therapy for social anxiety. An exception is the work presented in [8].

Singing, and especially singing within a choir, can have psychological benefits for individuals with chronic mental health issues, with research indicating a reduction in anxiety and an increase in positive emotions and relaxation, which translates to maintained and extended mental well being [3, 4]. However, empirical research in this area is in its infancy, with an acknowledged need for robust interdisciplinary studies [6]. From the technical perspective, no system exists that is capable of creating responsive 360 degrees experiences, enabling to sing with a virtual choir and interact with a virtual audience. However, research exists into the technologies and principles enabling such a system. Recent advances allow

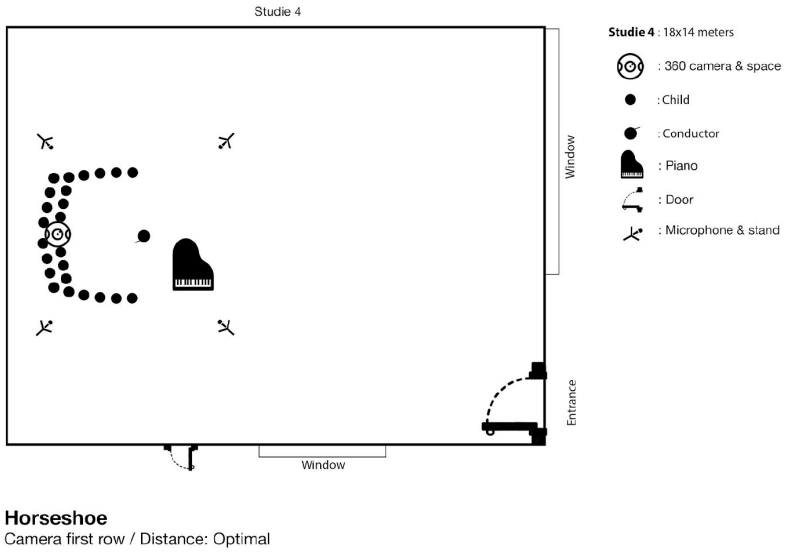


Fig. 4. The placement of camera, microphone and children while performing the songs.

for capturing and rendering 360 degrees videos, boosting research in so-called cinematic VR. This includes surround visual displays as well as spatial sound, particularly important to direct users attention to a specific locations of the screen [10]. Such solutions, however, have reduced interactivity, mostly limited to the possibility to turn the head in 360 degrees. The full interactive potential of VR has usually been adopted using 3D graphics computer rendered scenes. In this paper, we present the first prototype of a VR exposure therapy system, which combines 360 degrees video footage with the possibility of interaction and feedback. In the system, the users experience the venue of the original performance from the viewpoint of a singer, including visual and auditory feedback, and are able to hear themselves in that space as they sing live with the recorded performance. This system is intended to test the benefits of singing in a group in order to help coping with social anxiety. The system consists of a user singing while wearing an HMD.

2 Capturing the choir

During the Spring 2019, we had the possibility of recording the Danish National Chidren choir in one of their studios, shown in Figure 1. A subset of 25 children

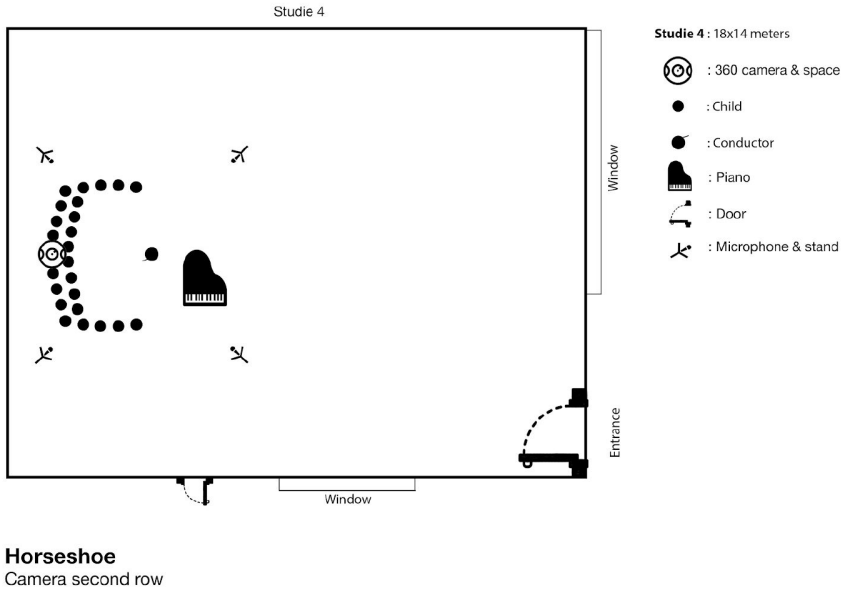


Fig. 5. The placement of camera, microphone and children while performing the songs.

from the choir were recorded while performing a warmup session and singing two songs from the repertoire of a day called *Skolernessangdag* [1]. This day is an event where several danish schools, for a total of around 200.000 children, learn different traditional danish songs, and sing them all together while the Danish National Children choir is streamed from the Danish radio concert hall. It is a very cozy event that promotes singing and being together.

We recorded the experience from different viewpoints, as shown in Figure 2,3 and 4,5,6. rated from inducing high anxiety towards inducing lower anxiety. Specifically, Figure 2 and 3 show the position of camera, microphones and children during vocal exercises. As can be seen from the figure, the children were placed in a circle. Two scenarios were captured, from a first scenario where the camera was very close to the children, to another scenario where the camera was far from the children. The two scenarios represented a situation with higher and

[1] The material from the day can be found here: https://skolernessangdag.dk/

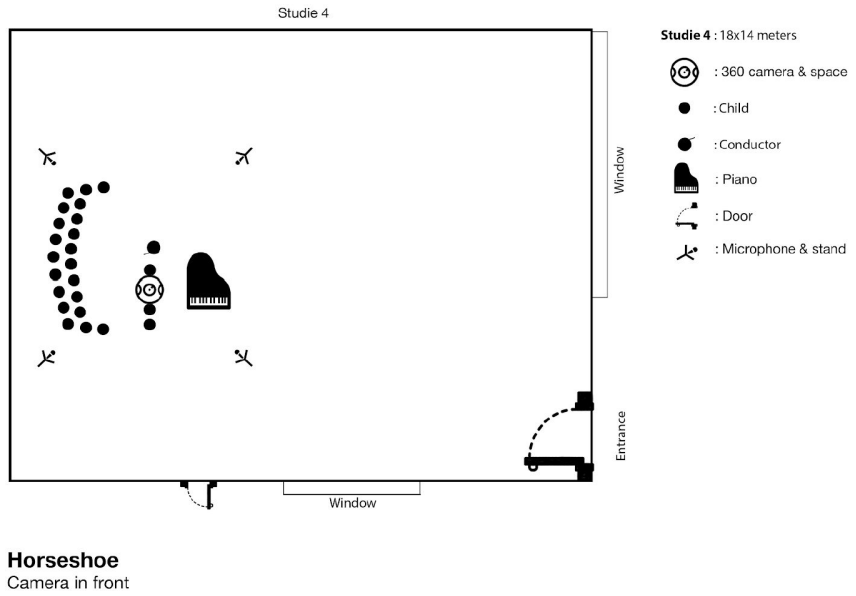


Fig. 6. The placement of camera, microphone and children while performing the songs.

lower social anxiety respectively. The recordings consisted of capturing the vocal exercises that the children usually perform at the beginning of a session.

Figure 4, 5 and 6 show the placement of the children while performing two songs. Specifically, the chosen songs were two famous danish songs named *Hvis jeg var en cirkushest* and *Tarzan Mamma mia*. Three situations were captured, from a situation inducing less anxiety (with the camera placed in the second row and with a large distance from the other children) to a situation inducing higher anxiety (where the camera is placed in the first row and very close to the other children).

Visuals were captured using an Insta 360 Pro Camera, while the sound was captured by the ambisonic microphone Ambeo by Sennheiser, as well as clip mics placed on the conductor.

3 Audio-Visual rendering

The virtual reality experience consisted of the captured footage from the Danish national children choir, together with 3D rendered relaxation rooms.

The captured videos were rendered in Unity and delivered using an Oculus head-mounted display.

Figure 9 shows the 3D rendered relaxation spaces. The goal of these relaxation spaces is to design virtual environments where the child can go to and take a break from the singing experience, in case it becomes too overwhelming. The spaces were designed after consultation with a psychologist, who described the need to take a break from the experience when it becomes too overwhelming.

From top to bottow, left to right, the spaces show a beach, and empty and a filled room, and then a forest. The spheres that can be seen in the figure are the different cinematic singing experiences. Such experiences can be grabbed by the user using the Touch controllers from Oculus, and are ordered according to the level of anxiety they induce.

The users singing is captured in real-time by a microphone, where is it processed by the reverberation characteristics of the desired simulated space.

Fig. 7. A screenshot from the captured footage during the warmup exercises.

4 Evaluation

A preliminary qualitative evaluation of the system was performed with two psychologists using the Cool Kids method and two interns who had suffered social anxiety and therefore used the Cool Kids method. The goal of the evaluation was to test how a psychologist can work with the child in VR through three phases

Fig. 8. A screenshot from the captured footage during the performance of one song, with the camera placed on the second row.

for exposure therapy (readiness, action, integration). The goal was to find design solutions for the phases. In the different phases there are also different design solutions to be found. The evaluation therefore tested: readiness phase, which belongs to a safe zone. This safe zone should help the children to relax and get ready for both VR and getting into a social scenario. For the psychologists they must be able to help the children relax and also choose which scenario to use played. This should lead on to a conversation topic on which functions the psychologist needs and how best to get the child to relax. Action phase, which belongs to the social scenarios. Here the different functionalities should be investigated. This space can be used by the psychologists to help them perform their work. For the children it is important to give them the commitment to participate and, if so, where their anxiety becomes too violent, they must be able to relax again. The sound in the action phase, which is being feedback to the player, must empower a child to sing further.

The integration phase, which belongs to safe zone or outside VR. Here the psychologist should be able to talk with the child about the experience they have been through. The psychologist intends to use VR to integrate the experience into a positive way for the child.

5 Method

The test was set up as a role-playing game where the psychologist and the child come in VR through the developed prototype. The evaluation was a qualitative evaluation based on feedback from the users. We first gave the psychologists the opportunity to talk about how they perform exposure therapy, in order to compare traditional methods with our VR solution. Since think out loud does

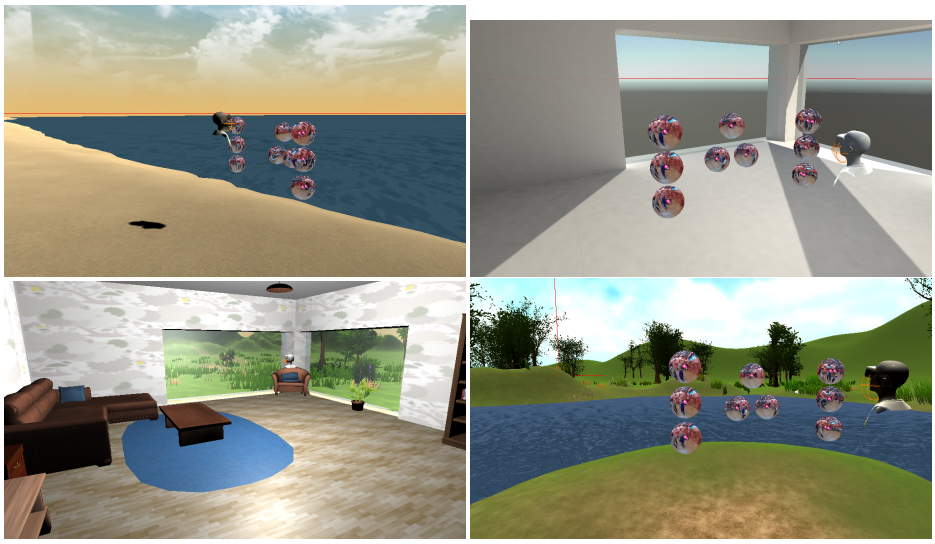


Fig. 9. The four different relaxation environments built for the application.

not work well with people who do not have experience with VR, because they may have trouble putting words on what they want, it is important to be able to ask questions while they are in VR, if they do not say anything for a while. This must be done without putting words into their mouths but hopefully starting a conversation about opportunities. The questions asked along the way were used as a springboard for a small conversation. By keeping it semi-structured, we could dive deeper into their thoughts and hopefully get them propose some solutions that would not otherwise have been said.

A joint discussion at the end, with all the test subjects, gave the participants the chance to talk about ideas, bouncing ideas between each other and gave us a chance to elaborate some of the things we ask about in the role-play scenario. Observations were taken along the way so that we have a basis for building questions in one joint discussion at the end. The observations looked at:

1. How could they use the digital space and integrate it into their daily process?
2. How will they interact with the system and with the child?
3. What do we expect them to do in relation to what they actually do?
4. What performance do they repeat?
5. What problems do they face?

The evaluation took place at the psychologists' clinic. A screenshot from the evaluation can be seen in Figure 10. In the screenshot, one of the two psychologists is wearing an head mounted display from Oculus while experiencing the simulation in his office, and providing think aloud feedback.

When asked to describe an ideal relaxation room, the first subject, a 19 years old girl with previous experience in anxiety, said she would like a homely environment with toys and books.

Fig. 10. One of the psychologists testing the application.

She also commented how the beginning of the experience was too loud and should be fixed. She suggested to implement fading in and out amplitude curves.

The second test user, a psychologist from the clinic, started by describing the importance of gradual exposure therapy. He then commented on the content of the waiting room. He suggested a place with more lively actions, for example the beach could have a seagull, or the forest could have birds and insects. He liked the implementation of several options, since each child is different and likes different forms of relaxation.

The third test user, an intern at Cool Kids, suggested to have the relaxation room look like the cafe' from the clinic, in order to create a virtual safe and comfortable environment.

She also found that the different levels on anxiety worked, e.g., she could perceive the difference in anxiety between performing from the second row to being in front of the other children where you feel placed on the spotline.

The last subject was the head of the clinic. He stressed how it is important to create a plan that goes through different steps in order to reach the child's goal. Improvisation and deviation are expected and the psychologist is prepared to do so. This plan also helps both parties for the different phases, as there is an expectations on both sides what the current session is about - control is a key component from both parties.

Regarding the readiness room, he was impressed and liked the metaphor of the beach, a zen-like places where there is no to little interruptions. Each recorded session was represented as a tiny ball that can be interacted with, grabbed and thrown. It is a way to normalize what is about to happen, and gives a trust/interaction with the child. Expressiveness is a must in order to fully read the emotional and physiological state of the child, and if its possible to transit to the next phase. Drawing, scales, or a dashboard to express themselves could be some possibilities. Visual expressive tool is a necessity, but should correlate to the surrounding environment. Simplistic seems to be a keyword for a readiness phase, where there is a possibility to reduce the incoming stimuli and find one self. Interaction is a necessity. Søren, the psychologist sees a lot of possibilities using just the first recordings over more sessions - inactive, active but only movement, active movement and singing. Then go to the next scene , where you are in the front row, do the same and go to the last position that is in front of the choir. Agrees there is an increasing social anxiety in the dataset A. Must be a possibility for the child to mark, point towards what can be a fearful factor. Time and time again it is necessary for the child to express their emotions and malaise. Complete control for the psychologist - play, pause, rewind, forwind etc. with the sessions, environment etc. Back and forth between scenes could also be a possibility. The child should not necessarily control the important, but should be able to express itself. Natural physiological body language.

Again as previously stated, it all depends on the child. Some might find it easier to integrate in VR, while others not so much. It is important to understand that this is a tool, and should not replace reality. Some evaluating tools in VR could be a fun way to try and integrate what we learned in the session.

6 Conclusion

In this paper, we presented a prototype and a primary evaluation of a VR experience that allows singing in a choir to prevent or help to cope with social anxiety. The purpose of this evaluation was to gather some information on how to further develop the application to make it more useful for professionals working with children suffering from social anxiety. During the evaluation, two clinicians and two interns at the clinicians who had suffered from social anxiety tried out the application and provided us with some valuable feedback.

The head of the clinic pinpointed the importance of enabling the opportunity to improvise during the VR exposure therapy, as he does during the traditional in-vivo exposure therapy sessions. One way to allow the clinicians to improvise during VR exposure therapy sessions is to let him pause, start or stop the ap-

plication, as well as giving him the ability to change the volume of the choir members or the choir instructors. Additionally, the ability to change scenes can be added to the application, giving the psychologist even more control of his patients experience.

The psychologists also pointed at the importance of interaction between the psychologist and the patient during exposure therapy sessions. During the traditional in-vivo exposure therapy sessions, the psychologist is present and can communicate to the patient about the anxiety-inducing experience, guiding him in how to cope with the situation. In VR, when the patient is immersed in another environment using the head-mounted display, communicating with the outside world becomes less efficient. Therefore, future iterations of the application can try to bridge the communication gap between the psychologist and the patient when using the VR exposure therapy application.

Via a microphone, the voice of the psychologist can be broadcasted in the virtual environment, while the volume can be adjusted so that it is loud and clear for the user. It is even possible to create a version of the application where the psychologist can log in to the VR exposure therapy together with the patient using a head-mounted display where they both can see, talk to and interact with each others using their avatars. Future studies should investigate how to design VR exposure therapy interventions that provide the ability to control the sessions while enabling them to improvise. Interaction schemes required to achieve this must be intuitive as the clinicians should not spend their cognitive capabilities on figuring out how to interact with the program and instead use it to help their patients.

7 Acknowledgments

We would like to thank Susanne Wendt and Clara Smedegaard, conductors at Danish National children choir, and the children for preparing a dedicated rehearsal session and allowing us to record their rehearsal. We would also like to thank the psychologists from Cool Kids who provided us with design suggestions as well as dedicated time to test our prototype.

References

1. Page Anderson, Barbara O Rothbaum, and Larry F Hodges. Virtual reality exposure in the treatment of social anxiety. *Cognitive and Behavioral Practice*, 10(3):240–247, 2003.
2. Gillian Butler. Exposure as a treatment for social phobia: Some instructive difficulties. *Behaviour Research and Therapy*, 23(6):651–657, 1985.
3. Stephen Clift, Grenville Hancox, Ian Morrison, Brbel Hess, Gunter Kreutz, and Don Stewart. Choral singing and psychological wellbeing: Quantitative and qualitative findings from english choirs in a cross-national survey. *Journal of Applied Arts & Health*, 1(1):19–34, 2010.

4. Genevieve A Dingle, Elyse Williams, Jolanda Jetten, and Jonathon Welch. Choir singing and creative writing enhance emotion regulation in adults with chronic mental health conditions. *British Journal of Clinical Psychology*, 56(4):443–457, 2017.
5. Edna B Foa and Michael J Kozak. Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, 99(1):20, 1986.
6. Mary L Gick. Singing, health and well-being: A health psychologists review. *Psychomusicology: Music, Mind and Brain*, 21(1-2):176, 2011.
7. Alessandra Gorini, Giuseppe Riva, Deacon, Gould, Kobak, Riva, Burdea, Brooks, Steuer, Riva, et al. Virtual reality in anxiety disorders: the past and the future. *Expert Review of Neurotherapeutics*, 8(2):215–233, 2008.
8. GAVIN Kearney, HELENA Daffern, LEWIS Thresh, HAROOM Omodudu, CALUM Armstrong, and JUDE Brereton. Design of an interactive virtual reality system for ensemble singing. In *Interactive Audio Systems Symposium, York, UK*, 2016.
9. Heidi J Lyneham, Vivana Wuthrich, and Ronald M Rapee. *Cool Kids: Child & Adolescent Anxiety Program Adaptation for Supported Bibliotherapy Therapist Manual*. Centre for Emotional Health, 2010.
10. Andrew MacQuarrie and Anthony Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*, pages 45–54. IEEE, 2017.
11. David Opriş, Sebastian Pintea, Azucena García-Palacios, Cristina Botella, Ştefan Szamosközi, and Daniel David. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. *Depression and anxiety*, 29(2):85–93, 2012.
12. Albert Skip Rizzo and Russell Shilling. Clinical virtual reality tools to advance the prevention, assessment, and treatment of ptsd. *European journal of psychotraumatology*, 8(sup5):1414560, 2017.

gravityZERO, an installation work for virtual environment

Suguru Goto[1], Satoru Higa[2], johnsmith[3], and Chihiro Suzuki[4]

[1] Department of Musical Creativity and the Environment, Tokyo University of the Arts Tokyo, Japan
goto.suguru@ms.geidai.ac.jp
[2] Backspace Productions Inc. Tokyo, Japan
satoruhiga@gmail.com
[3] 202 APwistaria Simotakaido 3-20-5 Suginami Tokyo, Japan
johnsmith13@iamas.ac.jp
[4] Graduate School of Art Nihon University Tokyo, Japan
mail@chihirosuzuki.com

Abstract. This paper reports the exposition of an artistic installation, gravityZERO, and its ongoing technical development. It consists of virtual sound, VR and robotic technologies in order to simulate the state of zero gravity. Audience members can experience a floating sensation within this virtual environment.

Keywords: Mechatronics and creative robotics, movement expression in avatar, artificial agents, virtual humans or robots, sound installation, VR/AR, projection mapping, 3D sound

1 Introduction

gravityZERO (zero gravity) is an installation that combines video, sound, and robotics. Translucent cubes are assembled at the venue. Images are holographically projected on the cube's surfaces, and speakers are placed in the cube's corners. A person is suspended from the ceiling and floats as if there is no gravity. Each rope can be freely moved in 3D space within the cube through motor operation. This project is based upon an idea that utilizes an interface closely related to a human body, and also pursues new possibilities of the performance of Augmented Reality and Projection Mapping.

2 Detailed Descriptions

In the last presentation, a mannequin was suspended instead of a person. By wearing a head-mounted display, the audience can experience the space as if they were weightless and floating through a binaural microphone and camera attached to the head of the mannequin. Since the previous presentation, safety has been confirmed, and people will be able to be suspended in the following project. This will allow participants to experience a zero-gravity state.

3 Technical Descriptions

Motors are used to suspend the participant. Three large motors are currently controlled by the computer. Eventually, there will be 4 units to make the movement of 3D space more free. Eight speakers are placed at each corner of the cube to virtually reproduce the 3D sound space. This is called Specialization or Ambisonic. The images are projected by video projectors on each side of the cube. The ceiling of the cube is open for the motor installation and the front side is open for the audience to see the work. Sound, video and motor movements are all synchronized.

3.1 The system as a whole

In this work, all software is connected via OSC centering on Ableton Live. The timeline is created based on the music data placed in Live, and all systems are synchronized. Due to the convenience of preparation, Touch Designer equips a simulator that reproduces the space-wise relationship between the image and the mannequin. It is possible to edit while confirming the movement in a computer.

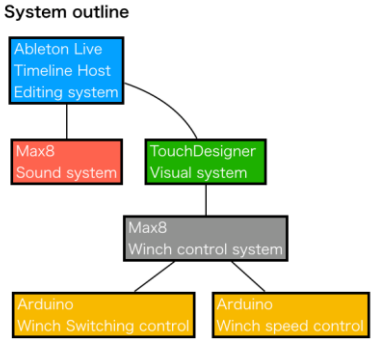


Fig. 1. The schematic diagram of an entire system

2.2 Robotics

The control system uses a hoist (RYOBI WI-62) and is produced by johnsmith and sheep. This robotic system is controlled by three winches which have a maximum load of 60 kg. The wires are connected to a stainless-steel triangle frame, and the height of the triangle's three points are changed according to the wire's pull. The mannequin's position is changed by the inclination of this frame connecting the three space coordinates. This alters the camera angle that relates to the eye position. As a result, it is possible to manipulate the viewpoint of the viewer through the camera attached to the mannequin.

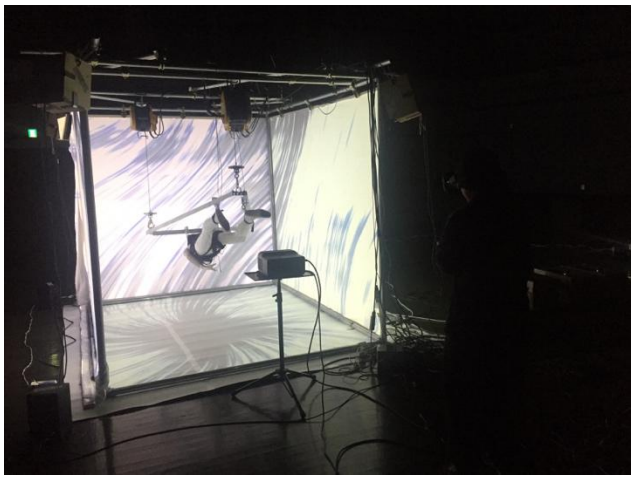


Fig. 2. The robotic system is realized with the winches, which are installed on the ceiling of cube.

The configuration of the device is controlled by Arduino and its Ethernet Shield, 12V control driver, which receives a signal from a computer via Ethernet. In this mechanism, raising and lowering is performed by a power relay, and acceleration/deceleration of speed is performed by a solid-state relay.

Fig. 3. Close view of motor and controller

Since WI-62 does not have a feedback value to measure the current position of the wire tip, the hook's current position resets the operating point of the automatic stop mechanism, which is already built within WI-62, and it calculates by adding the acceleration/deceleration at the time of the initial movement of the motor with the value in seconds. Based on this calculation, the rise and fall of the winch is managed by data in the timeline in Live.

As a safety measure, the winch wires are lowered vertically to prevent pulling each other. The length of wire corresponds to the height of each corner of triangle on the mannequin. This length is derived by the drive time of winding with the motor. However, irregular winding of wire may be caused by the winding diameter, in which the diameter may increase as the wires are winded more. This may be also be caused by an irregular motor drive voltage, and the continuous stop with short-time drive (such as PWM). The error is tolerated by the operation method, which returns to the reset state by the operation of continuing to pull back for 5 seconds for each wire at the possible timing.

The Max 8 control system monitors the winch condition and manages the winch control signal. OSC data from Live is sent via Touch Designer for the convenience of configuring a simulator in Touch Designer.

2.3 Visual System

The system in this work is produced by Satoru Higa. An image composed of a four-sided screen centered on a mannequin is organized in 3D space with Touch Designer. The camera is placed at the center of the generated visual as if it were floating in 3D space. This method was previously done by Higa's own software "VP3L" (Higa, 2006).

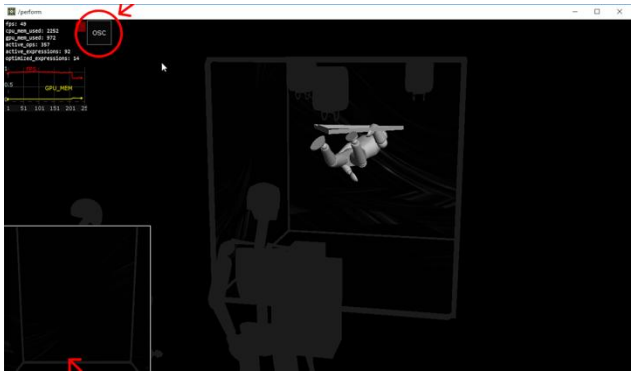


Fig. 4. The program for the images with TouchDesigner

The image is realized by four screens and four projectors arranged in a box-like space. The visual in 3D space generated by Touch Designer is cut out by the angle of view of four cameras arranged as a virtual camera in 3D space, resulting in an image projected on a four-sided screen. This is intended to be experienced in a box-like space, but the images are made to appear in real space by combining 4 surfaces which are originally cut out from one large 3D image like a window by each screen.

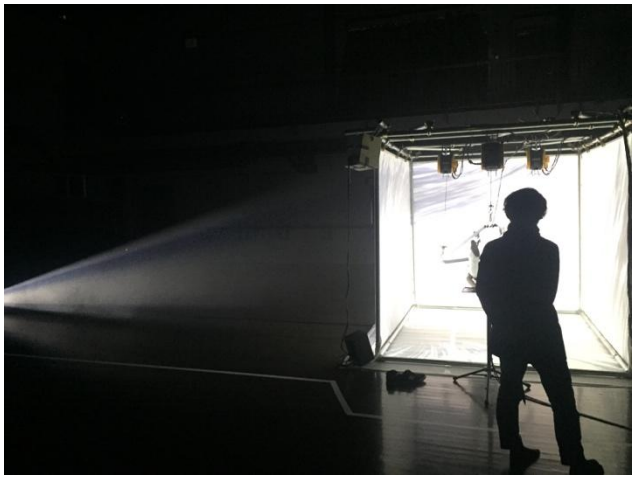


Fig. 5. The images on the cube, which are projected by 4 video projectors.

The black and white images appear according to the movement of the camera's viewpoint in 3D space. Therefore it creates a feeling of floating. The audio and visual elements interact. It also monitors the movement data of the winch, and incorporates effects such as changes in gravity that the body may receive, by moving the image.

2.4 3D Sound

Eight speakers are installed at each corner of the cube. These are connected directly with the Audio Interface, RME/Fireface UC. The sound moves among the 8 speakers as 3D sound of VR. The binaural microphones on the dummy head take this movement sound and relay it to the audience members for a thrilling effect. Due to its realistic nature, some audience members have experienced VR sickness.

For a technical description, the 3D sound is done with 8 simple panning in order for the sound to move freely in 3D space. This is done with Max 8.

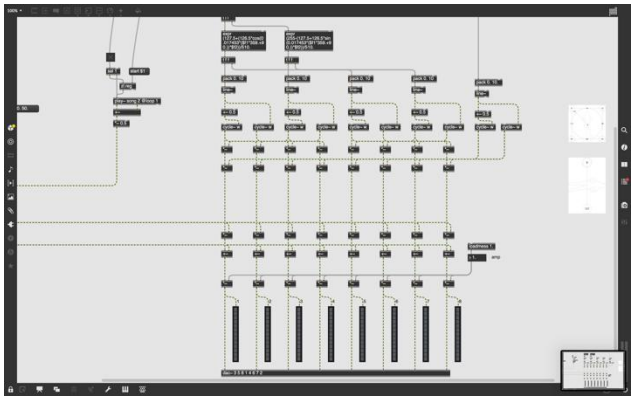


Fig. 6. The main patch of Spatialization.

The time line on Ableton Live manages the motors, sound and images. From M4L on Ableton Live, it sends the data via OSC in order to synchronize the Spatialization.

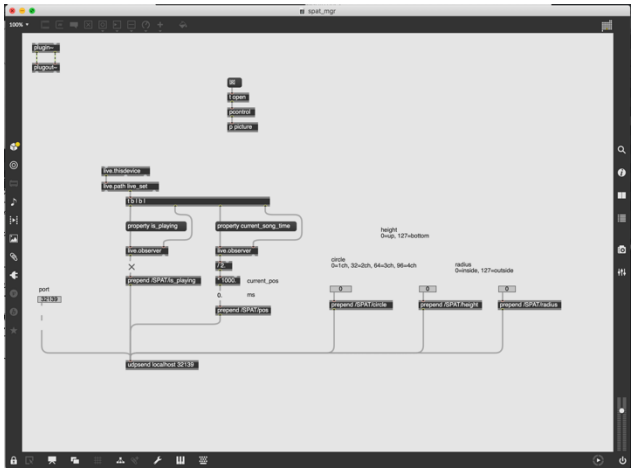


Fig. 7. The patch of OSC in M4L

2.5 gravityZERO as an Immersive Digital Environment

On the mannequin's head, a camera and CS-10EM: Binaural Microphones/Earphones are attached. The audience can listen to the sound with a

headphone, and see the images with HDM, HTC VIVE. This allows them to experience a sort of immersive digital environment (Immersion), which is a series of "Cave automatic virtual environment" (Cruz-Neira, et al., 1993) by Thomas DeFanti et al. In this work, three vectors may appeal to human perception by the surrounding image composed of four planes, physical control of the viewer's posture, and the positional change of the sound image.

Fig. 8. A camera and the binaural microphones are attached on the head of mannequin.

3 Results

Together, the abstract lighting and contouring lines in the image, the transitioning sounds and movement of the eyes and ears (camera and binaural microphone) of the viewer (mannequin) suspended in air, provides something that was not previously possible as a human experience.

This work does not provide reality, but presents a completely new experience, and suggests the possibility of a platform for a new viewing experience.

The work was presented at the exposition at Tokyo University of the Arts, Department of Musical Creativity and the Environment, Senju Campus, the 7th Hall on 2019.1.26. A large audience visited this exposition, and the work was well received. Most of the feedback was based on how the public observes the entire installation at first, especially how the mannequin floats in the air with the illusion of sound and images, and then how they were able to experience the state of zero gravity individually while wearing HMD and Binaural Headphones. People are

looking forward to the next exposition with the extended experience of being suspended in place of the mannequin.

4 Conclusions

gravityZERO is based on an artistic use of Mechatronics and creative robotics, especially with Movement expression in virtual humans and Sound Installation. To carry out the work, we applied the techniques of VR/AR, Projection Mapping, and 3D Sound. We look forward to developing gravityZERO to completion.

References

[1] F. Bevilacqua, R. Muller, and N. Schnell. MnM: a Max/MSP Mapping Toolbox. In NIME , Vancouver,
Canada, 2005.
[2] Frederic Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guedy, and Nicolas Rasamimanana: Continuous Realtime Gesture Following and Recognition, Computer Science (LNCS), Gesture in Rmbodied Communication and Human-Computer Interaction
[3] Rasamimanana, N.H., Bevilacqua, F.: E_ort-based analysis of bowing movements: evidence of anticipation e_ects. The Journal of New Music Research 37(4) (2009) pp. 339 – 351
[4] Rasamimanana, N.H., Kaiser, F., Bevilacqua, F.: Perspectives on gesture-sound relationships informed from acoustic instrument studies. Organised Sound 14(2) (2009) pp. 208 – 216
[5] Muller, R.: Human Motion Following using Hidden Markov Models. Master thesis, INSA Lyon, Laboratoire CREATIS (2004)
[6] Bevilacqua, F., Gu_edy, F., Schnell, N., Fl_ety, E., Leroy, N.: Wireless sensor interface and gesture-follower for music pedagogy. In: NIME '07: Proceedings of the 7th international conference on New interfaces for musical expression. (2007) pp. 124–129
[7] Bevilacqua, F.: Momentary notes on capturing gestures. In: (capturing intentions). Emio Greco/PC and the Amsterdam School for the Arts (2007)
[8] Schnell, N., Borghesi, R., Schwarz, D., Bevilacqua, F., Müller, R.: Ftm – complex data structures for max. In: International Computer Music Conference (ICMC). (2005)
[9] Bevilacqua, F., Muller, R., Schnell, N.: Mnm: a max/msp mapping toolbox. In: NIME '05: Proceedings of the 5th international conference on New interfaces for musical expression. (2005) pp. 85-88
[10] Viaud-Delmon, I., Bresson, J., Pachet, F., Bevilacqua, F., Roy, P., Warusfel, O.: Eartoy : interactions ludiques par l'audition. In: Journ_ees d'Informatique Musicale - JIM'07, Lyon, France (2007)
[11] Wanderley, M., Schnell, N., and Rovan, J.B.1998. "Escher - Modeling and Performing Composed Instruments in Real-Time." In Proc. IEEE SMC'98, pp. 1080-1084.
[12] Rovan, J., Wanderley, M., Dubnov, S., and Depalle, P. 1997. "Instrumental Gestural Mapping Strategies

as Expressivity Determinants in Computer Music Performance." In Proc. of the Kansei Workshop, Genova, pp. 68-73.
[13] Cruz-Neira, Carolina, Daniel J. Sandin, and Thomas A. DeFanti. "Surround-screen projection-based virtual reality: the design and implementation of the CAVE." Proceedings of the 20th annual conference on Computer graphics and interactive techniques. ACM, 1993.
[14] Walker John, Through the Looking Glass Beyond “User Interfaces”, The Art of Human-Computer Interface Design, Addison-Wesley Professional, 1990
[15] Sutherland, Ivan E. . "The Ultimate Display". Proceedings of IFIP Congress. pp. 506–508, 1965
Link:

http://gotolab.geidai.ac.jp/gravityzero/

Video:

https://vimeo.com/319681965

https://vimeo.com/319669023

Why People with a Cochlear Implant Listen to Music.

Jeremy Marozeau[1][0000−0002−4505−135X]

Hearing System Group, Department of Health Technology
Technical University of Denmark, Lyngby, Denmark
jemaroz@dtu.dk
www.hea.healthtech.dtu.dk

Abstract. The cochlear implant (CI) is the most successful neural prosthetic device in the market. It allows hundreds of thousands of people around the world to regain a sense of hearing. However, unlike a pair of glasses that can restore vision perfectly, the CI still has some shortcomings for non-speech sounds such as music and environmental sounds. Many studies have shown that most CI users have great difficulties perceiving pitch differences or recognizing simple melodies without words or rhythmical cues. Consequently, CI users report finding music less pleasant compared to their pre-deafness period. Despite this, many of those users do not entirely reject music, and it is not uncommon to see young CI users listening to music all day, or even playing an instrument. Listening to music is an experience that arises from more than the sum of the sensations induced by the basic elements of music: pitch, timbre and rhythm. Listening to music is a pleasant experience because it prompts high-level cognitive aspects such as emotional reactions, needs to dance, or the feeling of musical tension. Therefore, CI users still engaged in musical activities might experience some of these high-level features. In this paper, I will review recent studies on music perception in CI listeners and demonstrate that, although most CI users have difficulties with perceiving pitch, additional music cues such as tempo and dynamic range might contribute positively to their enjoyment of music.

Keywords: Cochlear Implant · Music and Deafness.

1 Introduction

The cochlear implant, CI, is a medical device that allows the direct stimulation of the auditory nerve fibers to restore some sense of hearing in severe to profoundly deaf people. Although it successfully helps many patients to perceive speech, the CI shows some limitations when reproducing musical signals [27]. Because of the inadequate frequency resolution created by the device, many CI users show scanty abilities in perceiving pitch, harmony, and timbre [25] [24]. However, despite these limitations, it is not uncommon to see young CI users engaged in musical activities. In this review, I will argue that, although their perception of pitch and harmony is limited, music is such a rich and complex signal that CI

users can rely on other high-level features to find some genuine enjoyment in musical activities.

2 The Cochlear Implant

In a healthy human cochlea, the sound is conveyed to the brain by the activation of the auditory nerve fibers connected to about 3.500 sensory receptors, called inner hair cells, located along of the cochlea. Unfortunately, damages to those inner hair cells, for example by exposure to loud music, are irreversible. This injury, commonly known as a sensorineural hearing loss, will prevent the acoustic wave from triggering any action potentials along the auditory nerve fibers. In that case, even the most powerful hearing aid will not be able to provide sufficient amplification to restore sound perception. A person suffering from such loss will be considered as candidate for a CI as this device can directly stimulate the auditory nerve and replace the function of the damaged inner hair cells.

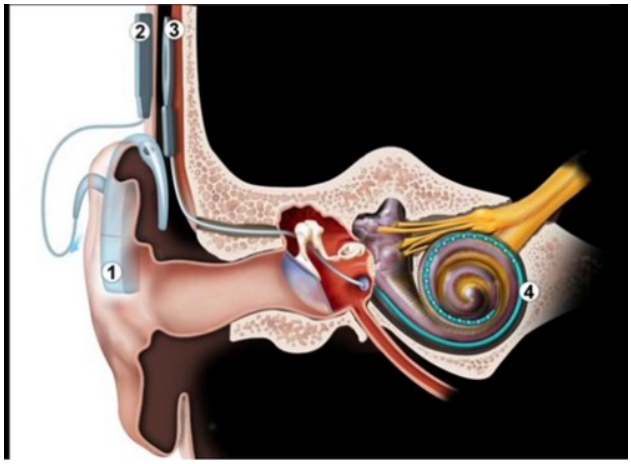


Fig. 1. Schema of a cochlear implant. The sound processor (1) captures a sound via its external microphone and converts it into a corresponding electrical signal. This signal is then transmitted through the skin by a radio-frequency from an external transmission coil (2) to the internal component (3). The electrical signal is then converted into an electric pulse that can generate an action potential in the auditory nerve (4) (Drawing by S. Blatrix for R. Pujol, www.cochlea.eu; used with permission of R. Pujol)

Figure 1 summarises the different parts and function of a cochlear implant. It is composed of an outer and an internal part. The outer part (#1) is a shelf that contains one or more microphones, the batteries, and the DSP chip that converts the acoustic signal into electric pulses based on a predefined sound coding strategy. Those pulses are transmitted as radio-frequency waves via a

wireless link antenna (#2) through the skin to the implant's receiver (#3). The antenna and the internal receiver are aligned by a pair of magnets. Finally, the pulses are delivered into the cochlea through a linear array of up to 22 electrodes (#4) and stimulate the auditory nerve directly, thus replacing the function of the hair cells that are lost or damaged in sensorineural deafness. The number of neurons activated will depend on the overall electric charge produced, the distance to the neurons, and the number of functional neurons.

3 The Perception of Music with a Cochlear Implant

Unfortunately, with the current technology, the CI cannot replace perfectly the role of the hair-cells. Therefore, many aspects of sound, important for music perception, will not be restored.

3.1 Pitch

The perception of pitch in normal-hearing listeners relies mainly on two possible coding mechanisms: the place and temporal coding. In the place coding, a pure tone will induce a maximum displacement on the basilar membrane at a specific location. Therefore, the brain will have an indication of the frequency of the sound by analyzing the position of the auditory nerve fibers activated. Given the high number of nerve fibers that homogenously innervate the cochlea, and the active behavior of the basilar membrane, such a cue could provide reliable information to extract a pitch. In the temporal coding, the overall activity of the auditory nerve fibers creates a temporal pattern that is linked to the frequency of the sound (at least up to about 1-4 kHz), giving a possible additional cue to form a pitch percept.

A cochlear implant has a limited number of electrode positions restricting the place coding dramatically. Furthermore, as the electrodes are located at a relatively large distance from the auditory nerve fibers, the current spread inside the cochlea will induce a substantial overlap between the region of neurons activated by each electrode. Additionally, most of the current coding strategies will use a fixed pulse rate that samples the envelope at a low frequency (typically around 500 to 900 Hz). Therefore, both natural pitch coding mechanisms will be only weakly represented in the current CI technology. It is therefore not surprising that studies on CI listeners found very weak abilities to discriminate pitch direction.

In most western music, the smallest pitch difference, the semitone, is about 6% change in fundamental frequency. However, Looi et al. [21] have shown that CI users need an average difference of at least 25% in fundamental frequency between notes, more than a minor third, to start to assess the direction of pitch change accurately. Such a low accuracy creates a clear challenge for CI listeners to follow a melody, in which the most common musical intervals between two notes are below a third [10]. Additionally, many studies using different tasks, such as pitch discrimination, melody recognition, or pitch change detection, have

shown that CI users perform, on average, significantly worse than normal-hearing listeners (e.g., [18] [13]).

3.2 Harmony

This weak ability to discriminate pitch will also result in an inability to perceive harmony and chord structure. Caldwell et al. [4] have asked ten CI users and 12 normal-hearing (NH) listeners to rate on a Likert scale from -5 (very unpleasant) to +5 (very pleasant) 36 piano pieces especially composed for that study. Those pieces were created based on 12 different melodies played with three possible types accompanying chords: 1) consonant triads, 2) minor seven chords, 3) dissonant minor seven with an augmented fifth. As expected, NH listeners judged consonant triads as the most pleasant, and dissonant chords as the least pleasant. On the other hand, CI users rated all three types of chords as equally pleasant. It is worth noting that despite having a very weak pitch percept, CI users report to enjoy every pieces. Recently, Knobloch et al. [14] showed that CI users rate major chords as more consonant than other types. These results suggest that consonance is somewhat accessible to at least some CI users. However, they are not able to differentiate an authentic cadence from a modified version in which the final tonic was replaced either by a transposed major chord or by a dissonant chord. Similar results were found with young implanted children [35].

3.3 Timbre

Along with the perception of pitch, timbre is also of paramount importance in music appreciation. It is a complex percept that can be divided into three main dimensions influenced by the temporal characteristic (impulsiveness), the spectral envelope (brightness) and spectro-temporal fluctuations [26]. Although the temporal variation of sound can be relatively well conveyed by the CI, it shows some limitations on spectral information [16] [17]. This results in a reduced accuracy of CI users to identify musical instruments [7] despite extensive training [6]. However, based purely on the attack time, CI users are still able to differentiate impulsive instruments, like the piano or the guitar, from non-impulsive instruments, like the flute or the violin [28] [11]. One might argue that the perception of timbre is not so important to enjoy music. Do we really need to be able to differentiate a trumpet from a saxophone to appreciate a tune from Miles Davis Quintet? Do we really need to listen to music with a high-fidelity sound system to experience some great joy out of it? While this might be true, it is worth noting that timbre is of paramount importance to segregate different instruments. Without that ability, a beautiful and sophisticated symphony may turn into a giant sonic mess.

3.4 Rhythm

It can be argued that rhythm is at the core of music. Fortunately, studies have shown that CI users can perform at nearly comparable levels as NH listeners

on simple rhythmical tasks such as pattern reproduction and discrimination [8], tempo discrimination [15], or dance in rhythm [30]. However, Jiam and Limb [12] argued that many of the studies on rhythm perception in CI users were based on relatively simple perception tasks, that differed from realworld music composed of multiple streams of many different notes. Such a dichotomy might reduce the ecological validity of those studies. In fact, the perception of tempo in music will depend on the ability to detect strong and weak beats. It relies on the ability to segregate different musical streams and on accurately identifying loudness differences. As both of those cues are affected by the CI sound coding scheme, it remains unclear how CI users can detect complex polyphonic rhythms or get an accurate sense of *groove*.

4 Why do CI users listen to music?

In the previous section, I have discussed many studies that showed that CI users have great difficulties in most of the building blocks of music. It should be natural then to assume that CI users would avoid music at any cost. In fact, Looi and She [22] have found with a survey on musical habits of over a hundred CI listeners that music was overall less enjoyable post-implantation. However, some CI listeners reported to enjoy music and to listen to it very often. Migirov et al. [29] showed that among 53 CI users tested, only 27% reported to never listen to music post-implantation. Interestingly, 30% of them still have a musical activity (playing an instrument or singing). A similar result was shown in a study by Brockmeier et al. [3], who found that 65% of their CI users regularly listen to music and 50% of them rate the quality via their CI as *pleasant*. One might wonder why do CI users still listen to music, given all the difficulties they have to perceive it. I will argue in the rest of the paper, that CI users can understand some higher-level features of music that allow them to find genuine enjoyment in musical activities.

4.1 The perception of musical emotion

Music is often considered as the language of emotion. It is, therefore, essential to understand the emotional content of a piece to enjoy it. Given that emotion is mostly conveyed in Western music by mode (major or minor) and tempo, CI users might have some difficulties to perceive it. Surprisingly different studies have shown that although they are not as precise as NH listeners, CI users are able to classify a piano piece accurately as *Happy* or *Sad* [33][1][2]. Given the lack of frequency resolution in the CI, it is unlikely that they only rely on the mode to perform this task. However, the question remains whether they are able to identify the emotional content based solely on pulsation.

For the sake of clarity, I will make a clear distinction in the rest of the document between the tempo, a musical instruction as noted by the composer on the music score, and the pulsation, a perceptual quality that can be defined as the heartbeat of the music or the pace at which someone will tap along the music.

If tempo is fixed and objective, pulsation can vary dramatically for a listener to another. For example, when a crowd starts clapping during a live performance, some people will tape every beat, some every downbeat, some every upbeat and some all over the place. However, after a few seconds, a consensus will appear, and most of the crowd will be more or less in sync. Therefore, although it can be argued that there is no single objective way to tap one music, one can extract a pulsation pace that will suit most people.

Vannson et al. [33] have asked 19 CI users to rate the intended emotion of 28 piano pieces specially composed to induce a specific feeling on a continuous slider labeled *Happy* and *Sad*. Among those 28 pieces, 14 were written in major mode, with tempi that varied from moderate to fast to convey happiness and 14 in minor, with tempi that ranged from slow to moderate to express sadness [34]. The published data suggest that the judgment of CI users can be modeled based solely on tempo.

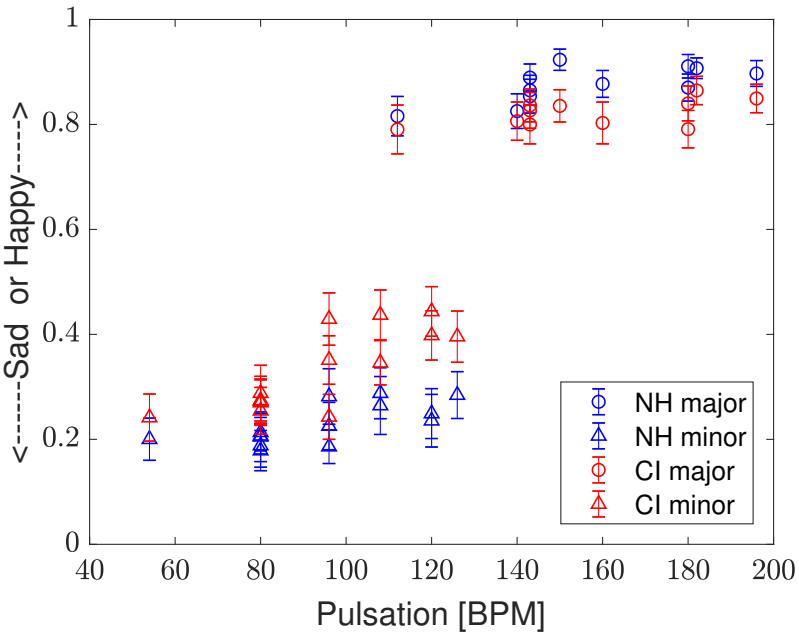


Fig. 2. Replot of the data from Vannson et al. [33]. The average emotional of rating 28 musical pieces is plotted as function of pulsation. The error bars represent the standard errors.

A recent re-analysis of this data suggests a slightly different conclusion. In the published analysis, the perceived pulsation of each piece was evaluated based on an online task in which NH listeners had to tap along with the original piano

pieces. However, it can be argued that CI listeners might judge slightly differently the pulsation of each piece as they have limited access to the pitch of each note. Therefore, to better estimate the pulsation of CI users, we have asked 10 NH listeners to tap along a modified version of each piece played on the congas. Although some variability was observed among the participants, for each piece at least one pulsation was produced by the majority of the participants.

Figure 2 replots the data as a function of pulsation. Regarding NH listeners (in blue), all the pieces in major mode (circle) are classified as *Happy* and all the pieces composed in minor mode as sad. Only averaged emotional rating of the minor pieces were moderately correlated with pulsation ($r(13) = 0.6988, p = 0.0054$). This result suggests that NH listeners relied primarily on mode. Surprisingly, a similar trend can be observed for CI users. All major pieces are judged as very happy (ranging from 0.79 to 0.86) and all the minor ones as sad (ranging from 0.24 to 0.44). Although both emotional ratings are correlated significantly with pulsation, the correlation was much stronger for the minor pieces ($r(13) = 0.5492, p = 0.0419$ for the major pieces, and $r(13) = 0.7997, p = 0.0006$ for the minor pieces). Although pulsation has an overall effect of the judgment of CI users, it cannot explain all the data. For example, a large difference can be observed between the minor pieces played at pulsation above 120 BPM, and the major ones played at a pulsation below 140 BPM. If the CI listeners had relied purely on pulsation, then those pieces should have been rated with similar emotion. Furthermore, it can be seen that a major piece with a pulsation of 110 BPM is judged as happier than minor pieces with similar or faster paces. Different hypotheses can be put forward to explain this result. First, the pulsations were derived from NH listeners tapping on the percussion version of the pieces to estimate the pulsation of CI users. This estimation might not be accurate. For example, the major piece at 110 BPM might be perceived as double by CI users. Second, it has been shown that CI users rate major chords as more consonant than minor chords [14]. Therefore, it cannot be excluded that CI users have access to the mode and will use it to form their judgments. Third, happy pieces have not only higher tempi, but they also have usually more syncopated rhythm, faster notes, and overall higher pitch. Therefore, CI users might have learned those subtle additional cues to extract the emotional content of the pieces accurately.

4.2 The perception of musical tension

An intense musical experience can often be linked to changes in musical intensity, or tension, in which the music gets gradually more dissonant and louder to reach a climax and then resolve into a smoothing and consonant moment that brings relief [31] [19]. Therefore, it can be hypothesized that to enjoy music, CI users must experience those different stages of musical intensity.

Tension rating can be modeled by a combination of tonal entropy, novelty, spectral centroid, and sound intensity [9]. **Tonal entropy** is related to the harmonic structure of music. In western music, tension is often created by subtle cues such as a dissonant chord, a delayed resolution or the resolution of a dominant

chord on the tonic [19] [20]. However, as CI users cannot discriminate between a consonant and dissonant chord progression [4] nor identify a cadence [14], it is unlikely that they will be able to use this cue to experience musical tension. **Novelty** relies on the capacity to memorize and identify pitch sequences. Given the weak ability of CI users to recognize well-known melody [5] it is unlikely that they will be able to rely on the novelty cue. **Spectral centroid** is related to the perception of the averaged energy content of the spectrum. Although as discussed sections 3.1 and 3.3, CI users cannot perceive fine differences in the spectral shape, they are still able to detect large variations of frequency, as needed to perceive vowels, and could perceive the spectral centroid in a similar way as NH listeners [23]. Finally, although the dynamic range of CI is limited to about 40 dB and often includes some heavy compressions, a monotonic relationship between the **sound intensity** and loudness can be achieved with CI users. Therefore, if CI users experience musical tension, they should rely mainly on loudness cues.

To test that hypothesis, Spangmose-Pedersen, Hjortkjær and Marozeau [32] have asked 9 NH and 9 CI listeners to rate the musical tension on a continuous analog slider of the Mozarts Piano Sonata No. 4 (K282) performed by an experienced pianist. In addition to the original piece, two modified versions were also tested in which 1) all the notes were altered, 2) the intensity of each keystroke was kept constant.

Figure 3 represents the average rating for both groups as a function of the musical measure. The piece can be divided into two repeating parts (A1, B1, A2, B2). The first 9 measures include the first part (A1) with a climax of maximal tension between the measures 6 and 7, and a release in 9. Measures 10 to 20 form the second part (B1), with two climaxes (12-13 and 15-16) and a long period of release (17-20). Then the piece is repeated with the same notes in parts A2 and B2 (20 til the end), but with a different interpretation of the musician. Surprisingly, the average rating pattern of CI users is highly correlated with the ratings of the NH users ($r(126) = 0.92385, p < 0.0001$). Additional statistical analysis (for more detail see [32]) outlined only few moments, during the climaxes of the section B1 and B2, in which the ratings of the CI users were significantly larger than those of NH listeners. This result suggests that CI users report a more intense experience of music. However, this interpretation should be taken with a caution, as CI listeners might have interpreted the task differently than the NH listeners.

In a follow-up condition, the listeners were presented with a modified version of the piece, where the pitch of each note was set to a random value. All other information, such as the timing and the velocity of the keystrokes were kept identical to the original piece. Results showed that removing tonal information had a much larger effect on the ratings of the NH listeners compared to the CI listeners. A third condition was tested in which the stimulus was a version of the piece where the velocity of the keystrokes was kept constant. This manipulation had a more pronounced effect on the ratings of the CI listeners than for the NH listeners. Overall, this experiment confirms that CI users can rate musical

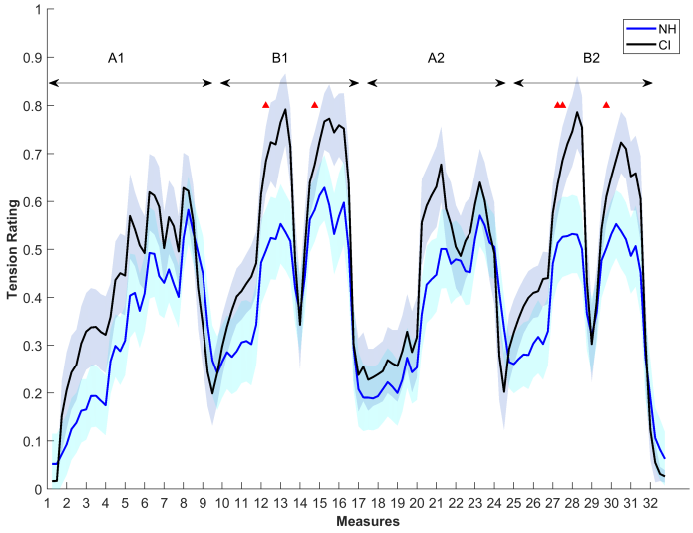


Fig. 3. Replot of the data from Spangmose-Pedersen et al. [32]. Average tension ratings of CI listeners (in black) and NH listeners (in blue) as a function of the musical measures. Horizontal double arrows outline the 4 parts of the piece (A1, B1, A2, B2). Shaded areas represent standard errors. Red triangles indicate the period in which the two ratings differ significantly.

tension in a very similar way as NH listeners, but that they rely mostly on intensity cues while NH listeners integrate many cues along with intensity such as pitch and harmony.

5 Conclusions

Although the cochlear implant (CI) can be highly successful in restoring speech perception in quiet, it still has some important shortcomings to convey the signals. Many studies have demonstrated that, overall, CI users perform below normal-hearing listeners in tasks relative to pitch, harmony, and timbre. However, they are still able to follow a rhythm change and a variation of intensity accurately. Based on this capacity, they can identify musical emotion, and follow changes in musical intensity. Although their overall percept of music is degraded, they seem to receive enough information to still be engaged in musical activities or simply enjoy listening to music.

6 Acknowledgments

Data on pulsation with NH listeners presented in section 4.1 was collected by Tanmayee Pathre during her final research project for her master degree. I would like to thank Niclas Janssen for his useful comments on an earlier version of the manuscript.

References

1. Emmanuelle Ambert-Dahan, Anne-Lise Giraud, Olivier Sterkers, and Severine Samson. Judgment of musical emotions after cochlear implantation in adults with progressive deafness. *Frontiers in Psychology*, 6, 3 2015.
2. SJ Brockmeier. Emotional response to music in combi 40/40+ users. *Cochlear Implants International*, 4(S1):25–26, 12 2003.
3. S.J. Brockmeier, P. Nopp, M. Vischer, W. Baumgartner, T. Stark, Schoen F., J. Mueller, T. Braunschweig, Busch R., M. Getto, W. Arnold, and D.J. Allum. Correlation of speech and music perception in postlingually deafCombi 40/40+ users. In TY Kubo and T Iwaki, editors, *Cochlear Implants - an Update*, page 599. Kugler Publications, 2002.
4. Meredith T. Caldwell, Patpong Jiradejvong, and Charles J Limb. Impaired Perception of Sensory Consonance and Dissonance in Cochlear Implant Users. *Otology & Neurotology*, 37(3):229–234, 3 2016.
5. Kate Gfeller, Christopher Turner, Maureen Mehr, George Woodworth, Robert Fearn, John Knutson, Shelley Witt, and Julie Stordahl. Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults. *Cochlear Implants International*, 3(1):29–53, 2002.
6. Kate Gfeller, Shelley Witt, Mary Adamek, Maureen Mehr, Jenny Rogers, Julie Stordahl, and Shelly Ringgenberg. Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. *Journal of the American Academy of Audiology*, 13(3):132–145, 2002.

7. Kate Gfeller, Shelley Witt, Maureen A. Mehr, George Woodworth, and John Knutson. Effects of Frequency, Instrumental Family, and Cochlear Implant Type on Timbre Recognition and Appraisal. *Annals of Otology, Rhinology & Laryngology*, 111(4):349–356, 4 2002.
8. Kate Gfeller, G Woodworth, D A Robin, S Witt, and J F Knutson. Perception of rhythmic and sequential pitch patterns by normally hearing adults and adult cochlear implant users. *Ear and Hearing*, 18(3):252–260, 1997.
9. J Hjortkjær. *Toward a Cognitive Theory of Musical Tension*. PhD thesis, Copenhagen University, 2011.
10. David Huron. Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception*, 19(1):1–64, 2001.
11. H Innes-Brown, J. Marozeau, C M Storey, and Peter J. Blamey. Tone, rhythm, and timbre perception in school-age children using cochlear implants and hearing Aids. *Journal of American Academy of Audiology*, 24(9):789–806, 2013.
12. Nicole T. Jiam and Charles J Limb. Rhythm processing in cochlear implant-mediated music perception. *Annals of the New York Academy of Sciences*, page nyas.14130, 6 2019.
13. R Kang, G L Nimmons, W Drennan, J Longnion, C Ruffin, K Nie, J H Won, T Worman, B Yueh, and J Rubinstein. Development and validation of the University of Washington Clinical Assessment of Music Perception test. *Ear and Hearing*, 30(4):411–418, 2009.
14. Marie Knobloch, Jesko L. Verhey, Michael Ziese, Marc Nitschmann, Christoph Arens, and Martin Böckmann-Barthel. Musical Harmony in Electric Hearing. *Music Perception: An Interdisciplinary Journal*, 36(1):40–52, 9 2018.
15. Ying-Yee Kong, Rachel Cruz, J Ackland Jones, and F G Zeng. Music perception with temporal cues in acoustic and electric hearing. *Ear and Hearing*, 25(2):173–185, 2004.
16. Ying-Yee Kong, A Mullangi, and J. Marozeau. Timbre and Speech Perception in Bimodal and Bilateral Cochlear-Implant Listeners. *Ear and Hearing*, 33(5):645–659, 2012.
17. Ying-Yee Kong, A Mullangi, J. Marozeau, and M Epstein. Temporal and Spectral Cues for Musical Timbre Perception in Electric Hearing. *Journal of Speech Language and Hearing Research*, 54:981–994, 2011.
18. J Laneau, M Moonen, and J Wouters. Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants. *Journal of the Acoustical Society of America*, 119(1):491–506, 2006.
19. F. Lerdahl. *Tonal pitch space*. Oxford University Press, New York, 2001.
20. Fred Lerdahl and C L Krumhansl. Modeling Tonal Tension. *Music Perception: An Interdisciplinary Journal*, 24(4):329–366, 4 2007.
21. Valerie Looi, H.J. McDermott, C. M. McKay, and L Hickson. Pitch discrimination and melody recognition by cochlear implant users. *International Congress Series*, 1273:197 200, 2004.
22. Valerie Looi and Jennifer She. Music perception of cochlear implant users: a questionnaire, and its implications for a music training program. *International Journal of Audiology*, 49(2):116–128, 2010.
23. Olivier Macherey and Alexia Delpierre. Perception of musical timbre by cochlear implant listeners: A multidimensional scaling study. *Ear and Hearing*, 34(4):426–436, 1 2013.
24. J. Marozeau and Wiebke Lamping. Timbre Perception with Cochlear Implants. In Siedenburg K., Saitis C., McAdams S., Popper A., and Fay R., editors, *Timbre:*

Acoustics, Perception, and Cognition, chapter 10, pages 273–293. Springer, Cham, 2019.
25. J. Marozeau, Ninia Simon, and Hamish Innes-brown. Cochlear implants can talk but cannot sing in tune. *Acoustics Australia*, 42(2):131–135, 2014.
26. S. McAdams, S Winsberg, S Donnadieu, G De Soete, and J Krimphoff. Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, 1995.
27. H.J. McDermott. Music perception with cochlear implants: a review. *Trends in Amplification*, 8(2):49–82, 2004.
28. H.J. McDermott and Valerie Looi. Perception of complex signals, including musical sounds, with cochlear implants. *International Congress Series*, 1273:201 204, 2004.
29. Lela Migirov, Jona Kronenberg, and Yael Henkin. Self-reported listening habits and enjoyment of music among adult cochlear implant recipients. *Annals of Otology, Rhinology and Laryngology*, 118(5):350–355, 2009.
30. Jessica Phillips-Silver, Petri Toiviainen, Nathalie Gosselin, Christine Turgeon, Franco Lepore, and Isabelle Peretz. Cochlear implant users move in time to the beat of drum music. *Hearing Research*, 321:25–34, 2015.
31. A. Schoenberg. *Style and idea.* St. Martins Press, New York, 1975.
32. Steffen Spangmose-Pedersen, J Hjortkjær, and J. Marozeau. Perception of Musical Tension in Cochlear Implant Listeners. *Frontiers in Auditory Cognitive Neuroscience*, 13, 2019.
33. N. Vannson, H. Innes-Brown, and J. Marozeau. Dichotic listening can improve perceived clarity of music in cochlear implant users. *Trends in Hearing*, 19, 2015.
34. Sandrine Vieillard, Isabelle Peretz, Nathalie Gosselin, Stphanie Khalfa, Lise Gagnon, and Bernard Bouchard. Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4):720–752, 2008.
35. Victoria Zimmer, Jesko L. Verhey, Michael Ziese, and Martin Böckmann-Barthel. Harmony Perception in Prelingually Deaf, Juvenile Cochlear Implant Users. *Frontiers in Neuroscience*, 13:466, 5 2019.

The 'Deaf listening'
Bodily qualities and modalities of musical perception for the Deaf

Sylvain Brétéché

PRISM (UMR 7061 - Aix-Marseille University/CNRS)

breteche@prism.cnrs.fr

Abstract. This paper focuses on 'Deaf listening' and more specifically on the involvement of the body in the Deaf musical experience. More precisely, it proposes to consider the sono-sensitive qualities of the body, from the Deaf experience, and by the investigation of a fundamental principle of the human experience: the 'corpaurality'. On the other hand, they are the natural modalities of Deaf hearing that will be discussed here, modalities which, detached from the ordinary aural perceptions, position the body at the center of the musical experience.

Keywords: Deaf people; Deaf listening; Deaf perception; Deaf musical experience; corpaurality; somato-sensitivity; embodiment.

1 Introduction

The Deaf[1] relationships to music are often thought to be altered, deteriorated or reduced because the ordinary consideration of deafness is that of the 'alteration'. Nevertheless, far from presenting only themselves as 'impaired beings', in their capacities of perception and representation of the world, Deaf take on this latter a singular view and, even more singular, give him an attentive 'ear', in tune with his sound manifestations.

[1] In accordance with the Deaf revendications, I write 'Deaf' with a capital D which, as specified by Charles Gaucher, "announces a quest for identity which falls into a very precise historicity and is stated in terms which seek to turn the deaf difference into a cultural particularity detached from the physical incapacity which stigmatizes it" [1, p. 17]. In this way, deafness proposes itself as a social and cultural group, where the constitutive dimensions of the community rely on Deaf specific features. In this article, I use to the term Deaf for name all the individuals who claims the Deaf identity and specificities – cultural, sensorial, social, linguistic,… – but also these specificities, which present themselves like particular qualities: Deaf qualities.
For information, around 466 million people worldwide have disabling hearing loss (over 5% of the world's population – disabling hearing loss refers to hearing loss greater than 40 decibel (World Health Organization estimations). Also, and according to the SIL International census and estimates (2019), there are 144 Sign Languages around the world. However, the number of native speakers of these Sign Languages remains difficult to establish formally, but can be estimated around 10 million (information available via www.ethnologue.com).

A priori paradoxical, the 'Deaf listening' is yet very real and upset our ordinary conceptions, highlighting that, if deafness is a human condition, it doesn't reveal 'alteration' but more especially 'otherness' [2]. Perceptive otherness or representational otherness, in the face of a reality that phenomenologically is not altered. Because the Deaf listening does not concern another sound world, but quite the contrary brings another approach of the common world, of this shared world that we define and fix from a 'normalized' point of view, audio-centered and finally limited to what our ears endeavor to say.

In this paper, I would like to focus on the Deaf listening and more specifically on the involvement of the body in the Deaf musical experience [3]. More precisely, it will be to consider the sono-sensitive qualities of the body, from the Deaf experience and by the investigation of a fundamental principle of the human experience: the 'corpaurality' [4]. On the other hand, those are the natural modalities of Deaf hearing that will be discussed here, modalities which, detached from the ordinary aural perceptions, position the body at the center of the musical experience.

2 Complexity of *corpaurality* principle

Essential base of sensoriality, the body presents itself as a sense vector and, faced with the sound reality, reveals a particular dimension, inherent in the human condition, what I call the *corpaurality*. Convergence of 'aurality' (what is perceived by the ear) and 'corporeality' (what is experienced by the body), two sensory modalities revealing the perceptible world, corpaurality designated the fundamental connection of the individual and the sound world: the body is anchored in the sensory worldn and the audible takes shape through it, form part of corporeality and reveals itself in an embodied way.

The 'hearing norm' that determines the ordinary delineation of the musical experience focuses primarily on the aural aspect of music. Indeed, "we must admit that when we play an instrument or listen to a disc, we use the sense that is socially intended for this purpose - hearing - and we consider most of the time that only the ear has a role to play in the listening function" [5, p. 54].

However, corpaurality as an essential principle of sound perception reveals that the music listening is naturally multi-sensory and confirms what is already known, that it is not only and exclusively located in the aural sphere of the perceptible world. As a sensory reality, the sound phenomenon fundamentally produces a material diffusion of mechanical vibrations, and for that "hearing is only one aspect of vibratory sense. The ear is not the exclusive receptor of sound vibrations, this function involves the whole body" [5, p. 56].

2.1 The vibratory sense

The vibratory sense informs about the sensory data perceived by the sense organs. The vibratory bodily sensitivity falls within *somesthesia*, that designates specificities of the body to perceive sensorial stimuli. The somatosensory system - or somesthetic system - concerns the sensitivity to stimuli perceived to whole the body, in

association or in addition to those directly in relation with the sense organs. Somatic sensations can thus give or supplement information on the environment. Unlike the sense organs, which concentrate their receptors in localized parts of the body (ears, eyes,...), the somesthetic system has receptors distributed over the entire body and positioned in the various layers that compose it: skin, bones, and musculo-tendinous or visceral levels.

The somato-sensitivity is based on specific sensory neurons - the mechanoreceptors, which perceive the stimuli and are attentive to mechanical transformations or deformations of the sensitive environment. Mechanoreceptors concern more specifically - and among others - the vibratory sensations; the human body is, in its entirety, sensitive to vibration frequencies, and this characteristically: "faced with the vibrations transmitted to the whole body, the human body behaves as a complex group [...]. At a given frequency, all or some parts of the body will react by amplifying the vibratory" [6, p. 45]. Thus, exposed to mechanical vibrations,

> the human body can be considered like an adaptable mechanical system, consisting of different entities connected to each other by springs and dampers that are the ligaments, muscles, intervertebral discs.
> When the body is exposed to vibrations, not all organs react in the same way. Each part of the body having its own resonant frequency. [7, p. 6]

The mechanical vibrations exploit a frequency field that ranges from about 1 to 3,000 Hz. Studies conducted on the resonance frequencies of the human body [6; 8] make it possible to determine a specific somato-sensitive organization, mainly developed at the level of the head, arms, bust, legs and spinal column. For an overall threshold of perception around 80 dB, the body naturally resonates at frequencies between 1 and 80 Hz, although the arms and hands are more sensitive - because resonating between 5 and 1,500 Hz. The reception spectrum is much lower than that of the human ear, which perceives frequencies between about 20 Hz and 20,000 Hz. However, we note that the somatic system is sensitive to infra sound - sound elements below 20 Hz - not perceived by the auditory system and considered inaudible.

The representation below shows a 'cartography' of the bodily resonances, on the basis of information specified by Johan Cardinaels [6]:

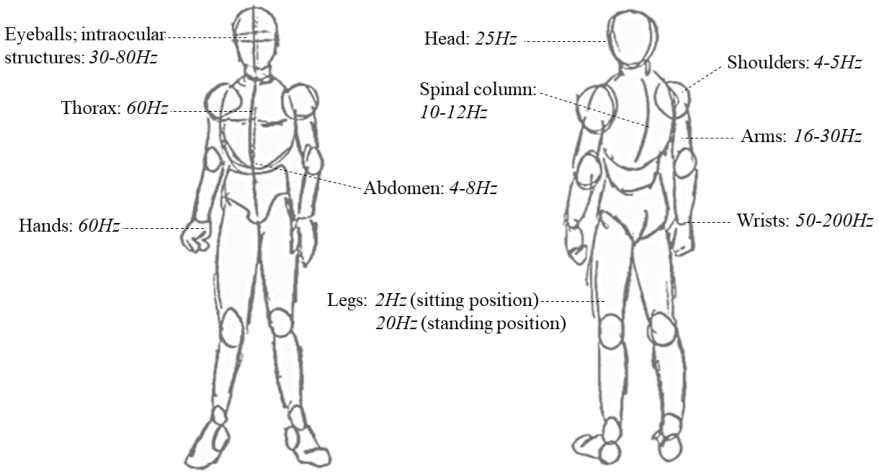


Fig. 1. 'Cartography' of bodily resonances [6; p. 44]

2.2 Modalities of the sono-sensitive body

The modes of bodily reception of mechanical vibrations are integrated into the somesthetic system, and can be categorized according to three distinct levels of sensitivity:

1) first, the exteroceptive sensitivity, which refers to the external perceptions, cutaneous, of mechanical variations. The body is directly in contact with the sensitive environment through the skin, which contains many sensory receptors – the exteroceptors – that react to air vibrations and are in charge of the vibrotactile perception. The cutaneous mechanoreceptors (or exteroceptors) reveal this sensitivity of the skin and are associated with three stimuli: pressure, touch, and vibration. Among the exteroceptors, two are more specifically engaged in the perception of mechanical vibrations: the Pacinian's corpuscles, which are sensitive to vibrations on a scale of 30 to 1,500 Hz - with an optimal sensitivity around 300 Hz, and Meissner's corpuscles, responsible for the 'fine touch' and particularly present in the dermis of hands, feet, lips and tongue, which are more sensitive to vibrations from 5 to 200 Hz.

2) The second level of somatic receptivity is the proprioceptive sensitivity, which brings together muscular, bony and tendon perceptions. Considered as the 'in-depth' reception, in opposition to the 'surface' perception represented by the exteroceptive sensitivity, the proprioceptive sensitivity allows the reception and transmission of sound vibrations through the whole body, through the musculoskeletal receptors. The bony reception of vibrations also concerns a particular aspect of the transmission of sound waves to the inner part of the

auditory system: the sound is not only transmitted by the sound waves to the middle ear, but also by bone conduction, the vibrations perceived by the body stimulating the inner ear via the cranium. Similarly, the auditory ossicles, which participate in the mechanical transformation of sound waves, can be stimulated by the cranium vibration.

3) The last level of the somatosensory system reveals the interoceptive sensitivity, which refers to visceral perceptions. The organs contained in the thoracic and abdominal cavities also contain numerous mechanoreceptors and the transmission of vibratory waves is carried out by the soft tissues contained in the body.

This specific organization of the somatic sensory system reveals the body's possibilities in the face of sound elements and materializes the corpaurality principle in its physiological reality. The somesthetic reception determines thus the faculty of the body to be sensitive to sound and concretizes the complexity of the corpaurality that formalizes the embodied qualities of listening.

3 The 'Deaf listening'

Corpaurality states this fundamental dimension of music reception and disrupts somehow the ordinary considerations of the musical experience. However, as the deaf percussionist Evelyn Glennie reminds us, "For some reason we tend to make a distinction between hearing a sound and feeling a vibration, in reality they are the same thing" [9]. To listen to music is feeling the mechanical vibrations of space, which concern the ear, but also and simultaneously the body. Within this context, Evelyn Glennie establishes a relevant connection with the Italian language, to highlight the fundamental link between listening and feeling:

> It is interesting to note that in the Italian language this distinction does not exist. The verb *sentire* means to hear and the same verb in the reflexive form *sentirsi* means to feel. Deafness does not mean that you can't hear, only that there is something wrong with the ears. Even someone who is totally deaf can still hear/feel sounds. [9]

Fundamentally, the sound world is felt and the otherness of the Deaf musical experience thus rests on this reality of feeling. Given their specificities, deafness situations reveal a singular apprehension of music that is fundamentally related to ordinary practices, but changes and reconsiders it outside the aural sphere. To be deaf is to feel the music vibrate, and as Danièle Pistone points out, "the hearing-impaired people themselves perceive the sound vibrations" [10, p. 69].

This Deaf musical otherness unveils the peculiarities of deafness as a human condition which, beyond revealing a hearing problem, suggests more precisely another modality of listening, 'denormed' and 'denormative' because fundamentally based on the materiality of sound reality. As Evelyn Glennie once again emphasizes, "to understand the nature of deafness, first one has to understand the nature of hearing" [9], and this can be extended by saying that, in a way, it is through deafness that could be to find out the deep nature of the hearing. Because corpaurality principle

determines the sono-sensitive qualities of the body in its anatomical constitution and its sensory possibilities, even though, in the ordinary musical experience, the listening remains focused on the ear as the privileged sensory organ. The hearing experience cannot escape the aural primacy, essential and natural medium of the musical experience, and it turns out that "only the deaf know what this means not to hear" [5, p. 56]. Therefore, the Deaf musical experience seems capable to restore a hidden but essential facet of hearing, which is in the first instance and in essence "a specialized form of touch" [9]. Basically, the listening aims to be "acoustical prehension" [11, p. 236], namely the grasp of the sound materiality in its vibratory consistency, which touches the ear but also the body in its entirety. The deaf experiences of music appear able to reveal this essential nature of hearing as grasping medium of audible reality. Listening to music is thus in the Deaf musical experience relocated from the ear to the body, which presents itself as the main base for understanding and expression of musicality.

4 The Deaf' sono-sensitivity

In the Deaf musical experience, the listening is revealed therefore primarily embodied and the modalities of the Deaf listening refer more specifically to the three levels of somato-sensitivity.

Firstly, we find a cutaneous perception, and indeed "the skin, as a sensory system, with all its aptitudes is, therefore, an essential organ for the deaf" [5, p. 58], and according to the study conducted by Maïté Le Moël, the most sensitive areas are "the fingers and the palm of the hand, the toes and the soles of the feet" [5, p. 57] - where we find, in particular, Meissner's corpuscles. As outlined the music therapist Alain Carré, "very often, deaf people make music 'bare ears' and often barefoot to have a complete vibratory perception" [12, p. 15]. The sense of the vibrations transmitted by the floor passes directly "by the soles of the feet on a massive scale" [9, p. 57], and the air perception of sounds is commonly experienced naturally, without hearing aids ("bare ears") in order to give sustained attention to the vibratory qualities of the music.

Even if elements can be perceived aurally, the Deaf prefer to listen to the music naturally, without artificial deformations. The abandonment of hearing aids during the musical practice is in line with the will of a 'natural practice' (the perceived sounds are not transformed) but also with the desire to avoid the amplification of sounds that are often unpleasant to the ear. Deaf people prefer to live the experience of music in the most natural manner possible, to keep a sound experience not deformed; and as pointed out Alain Carré:

> For the deaf person, the most pleasant perception will often be bare ears, natural since there is no deformation or discomfort of this amplification compared to the wearing of hearing aids, even if they are the best. But in terms of music, deaf musicians often prefer to work with their natural perception, especially since they rely heavily on the processing of vibratory information in itself, even if it does not produce an aural sensation. This vibration becomes relevant for the profoundly deaf person. [12, p. 15]

It seems obvious that "the deaf are very sensitive to tactile perception" [5, p. 57] and pay specific attention to structure-born vibrations of the objects and materials that surround them. Claire Paolacci specifies in particular that "the tactile listening is more immediate for the deaf" [13, p. 15] and that the aural perceptions are very often secondary, or less meaningful. The Deaf "know how subtle the answer given by the skin constantly caressed by the sound waves from various origins" [5, p. 57], and they are indeed able to identify a usual or everyday object from the vibrations it produces, without necessarily perceive their aural quality. As the deaf musician Maati El Hachimi explains: "the deaf person feels the vibrations in a car, knows if it is going faster or slower, if he/she is in a tractor or in a small car" [13, p. 49]. Hearing people are also endowed with this somato-sensitive capacity of identification of sound elements, but it is primarily their aural perceptions which are significant and meaning; by contrast, "for the people deprived of hearing, the sensory discrimination of the waves by the bodily perception can reach a subtlety that hearing people can hardly suspect" [14, p. 226].

In accordance with this, sono-sensitive experiences are essential in the daily life of Deaf, and as David Le Breton emphasizes,

> vibration sensitivity allows deaf people to gather information about their environment: recognize the voices of relatives, detect footsteps, identify musical moments, the passage of a car, the fall of an object. [15, p. 171]

In the instrumental practice, the sonic variations of instruments are also perceived cutaneously, mainly through the hands (directly in contact with the object). This perception of sound vibrations by touch is often sought by the deaf person during the musical experience, and the use of specific objects which materially restore and amplifying the air vibrations (balloon, rigid pipe, wooden crate) is common. Similarly, the Deaf put their hands on the speakers to feel the air vibration produce the sound emission. Cutaneous reception is thus presented as a fundamental sensory modality in the Deaf musical experience,

> the sense of touch can reach, through long learning and multiple experiences, a maximum sensitivity. It can give to hearing impaired people the pleasure of feeling their skin receive every sound wave. [5, p. 58]

This cutaneous perception is associated with a bony reception of sound vibrations, which is based mainly on a structure-borne perception of the acoustic elements: "the bones are actively touched and precisely vibrated by the sound waves which they receive and transmit through the limbs and the whole body" [5, p. 58]. The vibrations of the floor are perceived on the feet and "by the knees where they produce a rotational movement on the kneecaps" [5, p. 57]. Bone perception develops initially by the contact of a body part with a material element, primarily through the legs, which are in contact with the floor; and as a young deaf person reports: "*when I try to listen to music (without hearing aids) I feel by the feet the vibrations. It taps through the body*" [16, p. 44]. The structure-borne reception of the musical vibratory movement seems to begin with the feet and invests in the rest of the body, and according to Maïté Le Moël "the most sensitive bone areas are the spine, the pelvis (ilium, sacrum, coccyx), the shoulder girdle (clavicles and shoulder blades) and the thoracic cavity (sternum, ribs, vertebrae)" [5, p.58].

During the instrumental practice, the body is vibrated by the instrument, primarily on the arms, and the bone perception of acoustic variations is also efficient by air conduction because “the head is a bone region frequently vibrating with the acoustic waves and in particular the cranium, the frontal area and the lower jaw” [5, p. 58].

More complex to describe, the internal sensations induce by sound vibrations make it possible "to hypothesize that soft tissues are also good receptors of sound waves" [5, p. 58]. These contain many mechanoreceptors, and the descriptions offered by the deaf [4] of ‘resonances’ and ‘vibrations’ on the torso or ‘bubbling’ in the stomach indicate that the transmission of the vibratory waves is also performed via visceral conduction, which mainly concerns the thoracic and abdominal cavities.

This consideration of the sono-sensitivity of the Deaf musical experience confirms the reality of a specific bodily musical experience. The Deaf perceive and feel the music in and through their body, the latter receiving the sounds according to different perception modalities and develops a truly fine approach to sounds. The music therapist and anthropologist Alain Cabéro specifies in particular that the pitch is felt differently by the body:

> when the sound was low, they located it in the stomach, but also on the face, when the sound was rather high-pitched, they located it along the arms and the head.
> We also had as answers for the low sounds: knee and hand. The high-pitched is not always very well perceived in its delicacy, it was more by a feeling of pain, in the ears. [17, p. 67-68]

These elements show the complexity of the Deaf musical experience which, as the ordinary musical experience, reveals as many facets as it implies of individuals. Each deafness is different, and each experience of the music reports a singular lived experience, deeply embodied, but conveying different meaning values. Evelyn Glennie’s testimony shows this complexity of the relation to the sound and the possible diversity of the bodily experiences of the music that it induces:

> I spent a lot of time in my youth (with the help of my school percussion teacher Ron Forbes) refining my ability to detect vibrations. I would stand with my hands against the classroom wall while Ron played notes on the timpani (timpani produce a lot of vibrations). Eventually, I managed to distinguish the rough pitch of notes by associating where on my body I felt the sound with the sense of perfect pitch I had before losing my hearing. The low sounds I feel mainly in my legs and feet and high sounds might be particular places on my face, neck and chest. [9]

to make more concrete this somatic sensitivity of the Deaf, we can see the body areas stimulated during the musical experience, areas specified by the Deaf in a survey conducted in 2015 [4] and which precisely correspond to the ‘cartography’ of the bodily resonances:

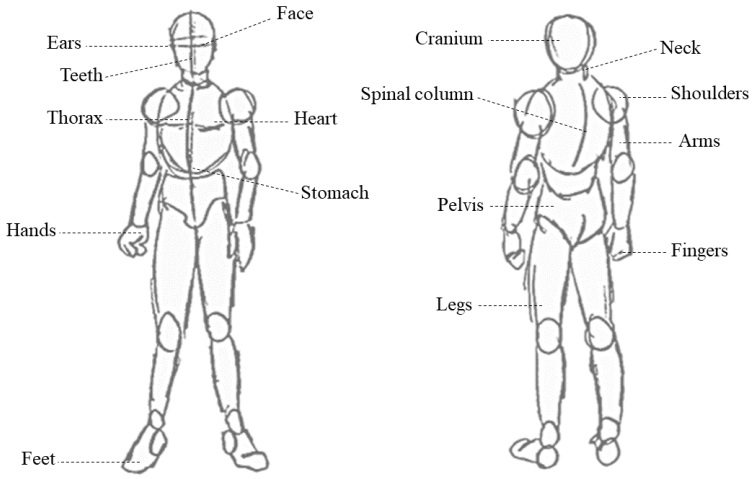


Fig. 2. Deaf's bodily perceptions [4, p. 588]

Despite the diversity of their musical experiences, we note that, for the Deaf, the body proposes itself as the privileged place of the musical realization. In this way, "regarding the direct experience of music since it is vibration, deafness, including profound, does not prevent sound contact, but displaces the privileged sense of the ear to the body" [14, p. 221]. The Deaf musical experience is fundamentally anchored in corpaurality principle and thus affirms, in the words of a young deaf person, that "*[we] can listen to music through the body and not through the ears. This is another way of listening. As if the body is the ear*" [18, p. 247]. Music is, therefore, fundamentally lived by the body and flourishes to become a sensitive reality revealing specific aesthetic values: "the whole body becomes 'the organ' of hearing [...]. By the vibrations that touch it, it replaces the ear" [18, p. 249-250].

*

Insofar as it gives fundamentally to feel, music imposes itself on the body; but even if "the link between the body and the music is complex and indissoluble" [19, p. 103], when the aural primacy is overcome, the musical experience finds within the body its space of realization. In this, to consider the Deaf musical practices makes it possible to go beyond the fundamental aurality of the music, by revealing the primordial role of the body. In other words, the Deaf otherness renders an experience of the music, detached from the ordinary conventions dependent to the ear performances, and seems to be able to reveal an unknown facet of the music, which makes body the essential organ of sound reception; and thus affirms that "hearing is not prerequisite to appreciating music" [20, p. 441].

References

1. Gaucher, C., Vibert, S. (ed.): Les sourds : aux origines d'une identité plurielle. Peter Lang, Bruxelles (2010)
2. Brétéché, S.: Through the prism of disability. For an experience of alterity: the being-deaf as figure of world. In: Dario, M. (dir.): The Role of Humanities in Contemporary Society: semiotics, culture, technologies, pp. 25-35. Kaunas University of Technology, Kaunas (2017)
3. Brétéché, S., Esclapez, C.: Music(s), Musicology and Science: Towards an Interscience Network. The Example of the Deaf Musical Experience. In: Aramaki, M., Davies, M, E.P.; Kronland-Martinet, R., Ystad, S. (eds.): Music Technology with Swing. 13th CMMR, pp. 637-657. Springer-Verlag, Heidelberg (2018)
4. Brétéché, S.: L'incarnation musicale. L'expérience musicale sourde. Thèse de doctorat en musicologie. Esclapez, C., et Vion-Dury, J., (dir.). Aix-Marseille Université (2015)
5. Le Moël, M.: L'univers musical de l'enfant sourd. In: Marsyas n°39/40, Dossiers Pédagogies et Handicaps. pp. 51-58 (1996)
6. Cardinaels, J.: Vibrations. Wolter Kluwer, Waterloo (2009)
7. CNAC (ed.): Bruit et vibrations. Fascicule n°112. Carl Heyman, Bruxelles (2006)
8. Chatillon, J. : Perception des infrasons. In: Acoustique et Techniques, n°67, pp. 4-10 (2011)
9. Glennie, E.: Hearing essay. https://www.evelyn.co.uk/ hearing-essay
10. Pistone, D.: Imaginaire et sens musical : des héritages aux réalisations. In : Grabócz, M. (ed.): Sens et signification en musique, pp. 35-49.Hermann éditeurs, Paris (2007)
11. Sayeux, A. S.: Le corps-oreille des musiques électroniques. Une approche anthropologique sensuelle. In: Communications, n°86, pp. 229-246 (2010)
12. Cité de la musique (ed.): Rencontre Musique et surdité. Cité de la Musique, Paris (2003) https://drop.philharmoniedeparis.fr/content/GPM/pdf/02Metiers04/Musique-et-surdite-cite-2003-06-24.pdf
13. Cité de la musique (ed.): Journée d'étude professionnelle Musique et surdité. Cité de la Musique, Paris (2005) www.citedelamusique.fr/pdf/handicap/260305_musique-et-surdite.pdf
14. Schmitt, P.: De la musique et des sourds. Approche ethnographique du rapport à la musique de jeunes sourds européens. In: Bachir-Loopuyt, T., Iglesias, S., Langenbruch, A., & Zur Nieden, G. (eds.), *Musik – Kontext – Wissenschaft. Interdisziplinäre Forschung zu Musik / Musiques – contextes – savoirs. Perspectives interdisciplinaires sur la musique*. pp. 221-233. Peter Lang, Frankfurt am Main (2012)
15. Le Breton, D.: La saveur du monde. Anthropologie des sens. Métailié, Paris (2006)
16. Cabéro, A.: La musique du silence, Éditions du Non Verbal/A.M.Bx, Parempuyre (2006)
17. Cabéro, A.: De l'ouïe à l'audition. Éditions du Non Verbal/A.M.Bx, Parempuyre (1998)
18. Cabéro, A.: Différent, différence et différends. Essai anthropologique sur les dissonances de la surdité mal-entendue. Thèse de doctorat de l'université de Bordeaux 2, mention ethnologie - option anthropologie sociale et culturelle, Traimond, B. (dir) (2009)
19. Csepregi, G.: La musique et le corps. Vladimir Jankélévitch sur l'art du piano. In: Csepregi, G. (dir.): Sagesse du corps, pp. 103-114. Édition du Scribe, Bruxelles (2001)
20. Loeffler, S.: Deaf music: embodying language and rhythm. In: Bauman, H-D.L.; Murray, J. (ed.): Deaf gain: raising the stakes for human diversity, pp.436-456. University of Minnesota Press, Minneapolis (2014)

Objective Evaluation of Ideal Time-Frequency Masking for Music Complexity Reduction in Cochlear Implants

Anil Nagathil and Rainer Martin *

Ruhr-Universität Bochum, Department of Electrical Engineering and Information Technology, Institute of Communication Acoustics, 44801 Bochum, Germany
anil@nagathil@rub.de, rainer.martin@rub.de

Abstract. Previous studies have shown that music becomes more preferable for CI listeners if the accompaniment in polyphonic music is moderately attenuated while the leading voice is fully retained. However, a recently proposed approach based on reduced-rank approximations of music signals suggests that reducing the spectral complexity of both leading voices and accompaniments can result in even higher benefits for CI users. In this paper we investigate this assumption in a simulated scenario. By applying ideal binary masks with different levels of attenuation to music signals, the relationship between a sole accompaniment attenuation and a reduction of higher-order leading voice harmonics is studied. An objective evaluation with an auditory-distortion measure and a music complexity prediction model for CI listeners predicts significant improvements. Hence, a benefit for CI users can be expected if the spectral complexity of leading voices is reduced in addition to an attenuation of the accompaniment.

Keywords: Hearing loss, music, cochlear implants, ideal binary mask

1 Introduction

A cochlear implant (CI) is a prosthetic device which is surgically implanted in the inner ear and bypasses the acoustic pathway of the auditory periphery by direct electrical stimulation of the auditory nerve. It can be used to alleviate effects of profound sensorineural hearing loss. While CI listeners reach an average speech intelligibility level of around 90% for full sentences within a two-years time span [23], most users are not yet satisfied with the perceived quality of music [10, 18]. This can be attributed to the high degree of current spread in the conductive perilymph of the cochlear and the entailing limitations in the number of electrodes usable in state-of-the-art CIs. As a consequence, frequency selectivity as compared to normal-hearing (NH) listeners is reduced, which distorts timbre and pitch cues considerably [6, 13, 12]. At the same time, CI users

* This work has been funded by the German Research Foundation (DFG), Collaborative Research Center 823, Subproject B3.

tend to prefer music played by solo instruments over ensemble music [11] and regularly structured music like popular and country music over more complex classical music [5].

Therefore, several studies have investigated strategies to improve the pleasantness of music for CI users by reducing the complexity of music signals [18]. For popular and country music it was shown that CI listeners rather prefer remixed versions of multi-track recordings with a moderate attenuation of accompanying instruments [2, 9]. Corresponding algorithms for performing a separation into harmonic and percussive elements or using supervised source separation methods have been proposed in [3, 19, 4]. In [15] spectral complexity reduction methods for music signals based on reduced-rank approximations were proposed and compared to a procedure based on supervised source separation and remixing. The methods were evaluated for classical chamber music in a listening test with NH listeners in combination with a procedure for simulating effects of reduced frequency selectivity and also with CI listeners [16]. For both listener groups the reduced-rank approximations, which attenuate low-energy harmonics of both the leading voice and the accompaniment, were significantly preferred over unprocessed music pieces and outperformed the remixing procedure. Therefore, it can be assumed that the perceived quality of music can be improved for CI users if the spectral complexity of the leading voice is also reduced besides an attenuation of the accompaniment. For an isolated melody this assumption was confirmed by attenuating higher-order harmonics using low-pass filters [17]. Although a reduction of harmonics generally leads to timbre distortions, this approach led to significant increases in pleasantness for CI listeners.

In order to confirm this hypothesis also for leading voices in the presence of a polyphonic accompaniment, in this paper we study the joint attenuation of accompaniments and reduction of higher-order harmonics of leading voices in a simulated scenario. To this end, we apply ideal binary masks (IBM) in the time-frequency domain, which are constructed by extracting fundamental frequency information of the leading voices from MIDI files. This allows to disentangle the effects obtained by processing the accompaniment and the leading voice, respectively. The performance is evaluated using an auditory-to-distortion ratio measure and a music complexity prediction model for CI listeners.

The remainder of this paper is organized as follows. Section 2 explains how IBMs are constructed to preserve certain harmonics of the leading voice. In Section 3 the evaluation setup is described. Results are presented and discussed in Section 4. Conclusions are drawn in Section 5.

2 Ideal Binary Masking

An ideal binary mask (IBM) is a multiplicative gain function with values of either one or zero, which is applied to individual time-frequency (TF) units of a signal. It has been applied for speech intelligibility improvement of noisy speech [22, 8] and for source separation in music signals [21]. In our context IBMs are defined in a slightly different way. We consider music signals with

a well-defined monophonic leading voice and a polyphonic accompaniment. To study the effects of the joint attenuation of accompaniments and higher-order leading voice harmonics, IBM gains are set to one for TF units of leading voice harmonics to be retained and to a pre-defined value between zero and one for the remaining TF units. To this end, we consider a signal $x(n)$ which is sampled at the sampling frequency f_s with n denoting the discrete-time index. A short-time spectral representation of the signal is obtained by computing the sliding window discrete Fourier transform (DFT)

$$X(k,\lambda) = \sum_{n=0}^{N-1} x(\lambda R + n)\, w(n)\, \exp\left(-j\frac{2\pi n k}{N}\right), \tag{1}$$

where k, λ, N, R, and $w(n)$ denote the frequency index, the frame index, the frame length, the frame shift, and a window function, respectively. Assuming that the temporal evolution of fundamental frequencies $f_0(\lambda)$ of the leading voice is given, we can identify the DFT frequency bins whose center frequencies are closest to the fundamental frequency and its harmonics, respectively, by

$$k_l(\lambda) = \arg\min_k |k\, f_s/N - (l+1)\, f_0(\lambda)| \tag{2}$$

with $k \in \{0, 1, \ldots, N/2\}$ and the harmonics index $l \in \{0, 1, 2, \ldots\}$. Then, an IBM can be constructed which preserves all TF units corresponding to the first m harmonics of the leading voice and attenuates all remaining TF units with a pre-defined gain. Hence, the mask is defined as

$$g(k,\lambda,m) = \begin{cases} 1 & k_0(\lambda), \ldots, k_m(\lambda) \\ 10^{\,0.05\, G_{\min}} & \text{otherwise,} \end{cases} \tag{3}$$

where $G_{\min}$ is a mask gain (MG) given in dB. The IBM is applied to the DFT spectrum yielding the processed DFT spectrum $Y(k,\lambda,m) = g(k,\lambda,m)\, X(k,\lambda)$. Computing an inverse DFT and applying the overlap-add procedure with a synthesis window $s(n)$ yields the processed signal $y(n)$.

An example of IBMs with $G_{\min} = -40$ dB and different numbers of retained leading voice harmonics with $f_0 = 391.995$ Hz (note G4) is shown in Figure 1. These IBMs are applied to a 0 dB mixture of a trumpet tone playing the note G4 and a piano chord playing the notes C4, E4, and C5. The resulting magnitude spectra are depicted for an exemplary frame in Figure 2. It can be seen that all frequency bins except for those corresponding to the leading voice harmonics to be retained are attenuated by 40 dB.

3 Evaluation

3.1 Database and Parameter Settings

The method described in the previous section was applied to extracts of 110 synthesized MIDI files of chamber music pieces (10 seconds long) which had

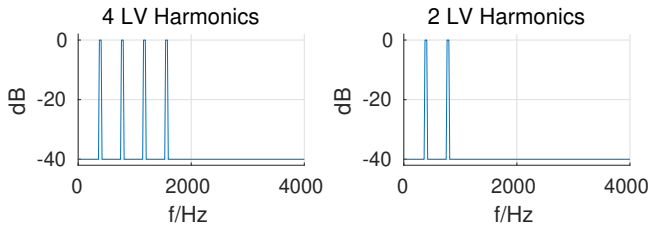


Fig. 1. Illustration of binary mask gain functions for the first four (left) and two (right) leading voice (LV) harmonics and a mask gain (MG) of $G_{\min} = -40$ dB.

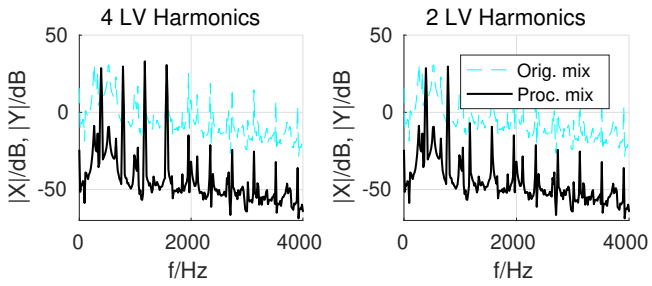


Fig. 2. Magnitude spectra for original and processed signal frames of an exemplary signal mixture (trumpet tone + piano chord) using the binary masks shown in Fig. 1.

already been used in [15]. Each of these pieces contained a well-defined monophonic leading voice played by a clarinet, a flute, an oboe, or a trumpet and an accompaniment played by a bassoon, a piano, or strings. The resulting leading melody and accompaniment signal waveforms were sampled at $f_s = 16$ kHz, converted to mono signals, and mixed at equal power.

For computing the DFT the frame length and shift were set to $N = 1024$ and $R = 512$, respectively. The analysis and synthesis window functions, $w(n)$ and $s(n)$, were both chosen as sqrt-Hann windows to satisfy a constant overlap-add constraint [7]. For constructing the IBMs we considered mask gains of $G_{\min} = \{-5, -10, -15, -20\}$ dB and the first one to four leading voice harmonics to be retained, i.e. $m = \{0, 1, 2, 3\}$. The fundamental frequencies f_0 were extracted from the original MIDI files. Since the DFT center frequencies are not matched to the geometric progression of fundamental frequencies in musical notes, spectral leakage occurs particularly at low frequencies. Therefore, the IBMs in (3) were adjusted to also pass through the contributions of the two frequency bins adjacent to $k_l(\lambda)$.

3.2 Evaluation Measures

In this work we used two instrumental measures which describe the amount of auditory distortions and predict music complexity as perceived by CI listeners. These two measure are outlined in the following.

In [1] a spectral smearing method was proposed to simulate effects of reduced frequency selectivity as a consequence of cochlear hearing loss. To obtain these spectrally smeared versions, a broadening of auditory filters is simulated. Based on this method, an auditory-distortion ratio (ADR)

$$\mathrm{ADR/dB} = 10\,\log_{10}\left(\frac{\sum_n \left[x(n)-\tilde{x}(n)\right]^2}{\sum_n \left[y(n)-\tilde{y}(n)\right]^2}\right) \tag{4}$$

was defined in [15] which quantifies the amount of auditory distortion resulting from spectral smearing. Here, $\tilde{x}(n)$ and $\tilde{y}(n)$ denote spectrally smeared versions of the original signal $x(n)$ and the processed signal $y(n)$, respectively. The numerator of the ADR measures the error between the original signal and its smeared version, whereas the denominator quantifies the deviation between the processed signal and its spectrally smeared counterpart. By definition, an ADR of 0 dB describes no change in the amount of auditory distortion and positive values indicate a reduction of auditory distortions for the processed case.

A music complexity prediction model for CI listeners was developed in [14]. In this work CI listeners provided music complexity ratings on a bimodal scale in the continuous range between -3 and $+3$ for a set of purely instrumental, non-percussive music pieces. Here, -3 describes the *extremely complex* case, $+3$ indicates the case of *extremely simple*, and 0 denotes a neutral level. The pieces were balanced in terms of subjective complexity as perceived by NH listeners. The complexity model predicts the median complexity ratings across all CI users using a principal component regression model which was trained with signal-based features. These features describe the amount of high-frequency energy in the signal, the frequency region with the highest concentration of spectral energy, the spectral bandwidth, and the degree of dissonance between pairs of spectral peaks.

Note, that both the ADR and the music complexity measure show a high and significant degree of correlation with CI listener preference scores obtained for processed versions of the music database used in this paper [18, 14]. This makes them suitable for an objective evaluation of CI-related music processing schemes.

To evaluate the statistical significance of the results, an analysis of variance (ANOVA) was performed for each measure followed by post-hoc tests with Bonferroni corrections of the significance level ($\alpha = 0.5$).

4 Results

4.1 Auditory-Distortion Ratio

Figure 3 shows the ADR results obtained for the processed signals. If all leading voice harmonics are retained (ALL), the ADR measure shows improvements of up to 1.4 dB after applying an MG between $G_{\min} = -5$ dB and $G_{\min} = -20$ dB. This corresponds to the benefit gained by performing a sole attenuation of the

accompaniment and therefore confirms the results of [2, 9, 19]. Retaining only the first harmonic of the leading voice and attenuating all other frequency bins yields a maximal improvement of 1.5 dB for $G_{\min} = -20$ dB. For all conditions, except for $G_{\min} = -20$ dB, retaining only the first leading voice harmonic yields a significant improvement over the case of preserving all leading voice harmonics. Note, that for decreasing values of $G_{\min}$ the ADR measure becomes more dependent on the fundamental frequency of the leading voice. This is reflected by an increase of variance in the ADR measure.

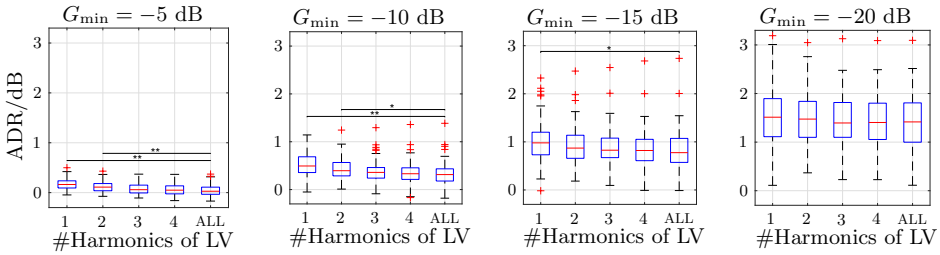


Fig. 3. ADR shown for different MG conditions and different number of retained leading voice (LV) harmonics. Significant ($p < 0.05$) and highly-significant ($p < 0.001$) differences are indicated by brackets with single and double asterisk symbols, respectively.

4.2 Music Complexity Prediction Model

Figure 4 shows the music complexity predictions obtained for the processed signals. The values '1', '2', and '3' describe the levels *slightly*, *quite*, and *extremely* towards each direction and '0' denotes a neutral level. For the unprocessed case (UN) a median prediction of -0.4 is obtained indicating that the music pieces are perceived as slightly complex. Attenuating the accompaniment whilst retaining all leading voice harmonics (ALL) yields a significant improvement on the scale with median results close to the neutral level across all MG conditions. Similarly as for the ADR measure, this improvement is in line with the results by [2, 9, 19]. If the higher-order harmonics of the leading voice are attenuated in addition, a further improvement is predicted across all MG conditions. This improvement is significant compared to the unprocessed case for all considered numbers of retained leading voice harmonics. Compared to the case *ALL*, a further attenuation of leading voice harmonics yields significant improvements if only the first one or two leading voice harmonics are retained (depending on the MG condition). The highest benefit is predicted for $G_{\min} = -20$ dB indicating that music is perceived as slightly simple after a reduction of the leading voice harmonics.

5 Conclusions

We investigated if reducing the spectral complexity of leading voices in polyphonic music signals in addition to an attenuation of the accompaniment can

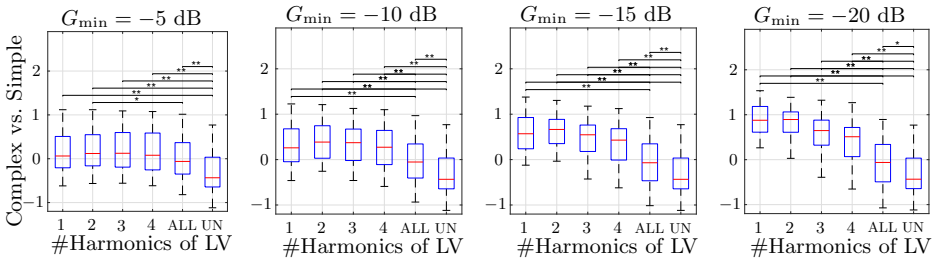


Fig. 4. Predictions of music complexity for different MG conditions and different numbers of retained leading voice (LV) harmonics. Significant ($p < 0.05$) and highly-significant ($p < 0.001$) differences are indicated by single and double asterisk symbols, respectively.

result in a benefit for CI listeners. To this end, we applied ideal binary masks to classical chamber music pieces with different mask gains and evaluated the quality of the processed music signals with instrumental measures. The results predict significant reductions in terms of auditory distortions and perceived complexity compared to a sole attenuation of the accompaniment. This explains why reduced-rank approximations, which attenuate low-variance contributions of both leading voices and accompaniments, were found to outperform approaches based on source separation and remixing [15]. Given the results of this study, it can be expected that binary time-frequency masking will result in significant benefits for CI users. Future research directions should therefore consider a blind construction of binary masks with adjustable gains by using estimates of pre-dominant melody contours [20].

References

1. Baer, T., Moore, B.C.: Effects of Spectral Smearing on the Intelligibility of Sentences in Noise. J. Acoust. Soc. Am. (JASA) **94**(3), 1229–1241 (1993)
2. Buyens, W., van Dijk, B., Moonen, M., Wouters, J.: Music Mixing Preferences of Cochlear Implant Recipients: A Pilot Study. Int. J. Audiology **53**(5), 294–301 (2014)
3. Buyens, W., van Dijk, B., Wouters, J., Moonen, M.: A Stereo Music Preprocessing scheme for Cochlear Implant Users. IEEE Trans. Biomedical Engineering **62**(10), 2434–2442 (2015)
4. Gajecki, T., Nogueira, W.: Deep Learning Models to Remix Music for Cochlear Implant Users. J. Acoust. Soc. Am. (JASA) **143**(6), 3602–3615 (2018)
5. Gfeller, K., Christ, A., John, K., Witt, S., Mehr, M.: The Effects of Familiarity and Complexity on Appraisal of Complex Songs by Cochlear Implant Recipients and Normal Hearing Adults. J. Music Therapy **40**(2), 78–112 (2003)
6. Gfeller, K., Turner, C., Mehr, M., Woodworth, G., Fearn, R., Knutson, J.F., Witt, S., Stordahl, J.: Recognition of Familiar Melodies by Adult Cochlear Implant Recipients and Normal-hearing Adults. Cochlear Implants International **3**(1), 29–53 (2002)

7. Goodwin, M.: The STFT, Sinusoidal Models, and Speech Modification. In: Benesty, J., Sondhi, M.M., Huang, Y. (eds.) Springer Handbook of Speech Processing, pp. 231–232. Springer (2008)
8. Hu, Y., Loizou, P.C.: Techniques for Estimating the Ideal Binary Mask. In: Proc. Intern. Workshop on Acoustic, Echo and Noise Control (IWAENC). pp. 154–157 (2008)
9. Kohlberg, G.D., Mancuso, D.M., Chari, D.A., Lalwani, A.K.: Music Engineering As a Novel Strategy for Enhancing Music Enjoyment in the Cochlear Implant Recipient. Behavioural Neurology **2015** (2015)
10. Limb, C.J., Roy, A.T.: Technological, Biological, and Acoustical Constraints to Music Perception in Cochlear Implant Users. Hearing Research **308**, 13–26 (2014)
11. Looi, V., McDermott, H., McKay, C., Hickson, L.: Comparisons of Quality Ratings for Music by Cochlear Implant and Hearing Aid Users. Ear and Hearing **28**(2), 59S–61S (2007)
12. Looi, V., McDermott, H., McKay, C., Hickson, L.: Pitch Discrimination and Melody Recognition by Cochlear Implant Users. International Congress Series **1273**, 197–200 (2004)
13. McDermott, H.J.: Music Perception with Cochlear Implants: A Review. Trends in Amplification **8**(2), 49–82 (2004)
14. Nagathil, A., Schlattmann, J.W., Neumann, K., Martin, R.: Music Complexity Prediction for Cochlear Implant Listeners Based on a Feature-based Linear Regression Model. J. Acoust. Soc. Am. (JASA) **144**(1), 1–10 (2018)
15. Nagathil, A., Weihs, C., Martin, R.: Spectral Complexity Reduction of Music Signals for Mitigating Effects of Cochlear Hearing Loss. IEEE/ACM Trans. Audio, Speech, and Language Process. **24**(3), 445–458 (2016)
16. Nagathil, A., Weihs, C., Neumann, K., Martin, R.: Spectral Complexity Reduction of Music Signals Based on Frequency-domain Reduced-rank Approximations: An Evaluation with Cochlear Implant Listeners. J. Acoust. Soc. Am. (JASA) **142**(3), 1219–1228 (2017)
17. Nemer, J.S., Kohlberg, G.D., Mancuso, D.M., Griffin, B.M., Certo, M.V., Chen, S.Y., Chun, M.B., Spitzer, J.B., Lalwani, A.K.: Reduction of the Harmonic Series Influences Musical Enjoyment With Cochlear Implants. Otology & Neurotology **38**(1), 31–37 (2017)
18. Nogueira, W., Nagathil, A., Martin, R.: Making Music More Accessible for Cochlear Implant Listeners: Recent Developments. IEEE Signal Process. Mag. **36**(1), 115–127 (2019)
19. Pons, J., Janer, J., Rode, T., Nogueira, W.: Remixing Music Using Source Separation Algorithms to Improve the Musical Experience of Cochlear Implant Users. J. Acoust. Soc. Am. (JASA) **140**(6), 4338–4349 (2016)
20. Salomon, J., Gómez, E., Ellis, D.P., Richard, G.: Melody Extraction from Polyphonic Music Signals. IEEE Signal Process. Mag. **31**(2), 118–134 (2014)
21. Virtanen, T., Mesaros, A., Ryynänen, M.: Combining Pitch-based Inference and Non-negative Spectrogram Factorization in Separating Vocals from Polyphonic Music. In: Proc. SAPA@INTERSPEECH. pp. 17–22 (2008)
22. Wang, D.: On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis. In: Speech separation by humans and machines, pp. 181–197. Springer (2005)
23. Wilson, B.S., Dorman, M.F.: Cochlear Implants: A Remarkable Past and a Brilliant Future. Hearing Research **242**, 3–21 (2008)

Evaluation of new music compositions in live concerts by cochlear implant users and normal hearing listeners

Waldo Nogueira[1]

[1] Hearing4all, Dept. of Otolaryngology, Medical University Hannover, Cluster of Excellence Hearing4all, Hannover, Germany
Nogueiravazquez.waldo@mh-hannover.de

Abstract. Cochlear implants (CI) have become very successful in restoring hearing abilities of profoundly hearing-impaired people. However, music perception remains generally poor for CI users. Typically, music perception with CIs is investigated under laboratory conditions. The present work investigates music perception in CI users in a real live concert. With this purpose a group of CI users, musicians and researchers worked together to compose music pieces. These compositions were interpreted in a concert where the appreciation of the music was evaluated by means of a questionnaire to compare the different audience listening groups (CI users and normal hearing listeners). In total, 253 people registered for the musIC 1.0 concert, 133 filled the questionnaire from which 37 were CI users and 8 were hearing aid users. As expected, the group of CI users rated melody, timbre and rhythm perception significantly lower than the group of normal hearing listeners, however no significant group differences were found on measures of enjoyment and cognitive aspects such as interest, emotion and understanding. In general, CI users preferred simple music with clear melodies.

Keywords: music, cochlear implant, music appreciation, music evaluation, hearing loss.

1 Introduction

Cochlear implants (CIs) are medical devices that can successfully restore speech communication, especially in environments with low or absent background noise. However, speech understanding in noisy environments and music perception with these devices are still severely limited. Limitations in music perception by CI users have been mostly investigated under laboratory conditions. The present work investigates music perception in CI users in live concerts.

It has been shown that CI users experience difficulties in recognizing melodies and distinguishing between different instruments in comparison to normal-hearing listeners [1], [2]. These limitations are related to the bottleneck created between the CI electrodes inserted in the cochlea's scala tympani and the auditory nerve fibers. The low number of electrodes, the broad excitation patterns and the channel interactions created when current is delivered through the CI electrodes as well as the

limited transmission of temporal fine structure (TFS) constitute the bottleneck. As a result, many CI users often report that music is perceived as too complex and not pleasant. For example, it has been shown that CI users prefer music with predominant vocals or based on single instruments over ensemble or orchestra music and that regularly structured pop and country music are favored over classical music [1], [3], [8].

Researchers and musicians have developed innovative methods to make music more accessible to CI users beyond signal processing and sound coding developments (e.g. [6], [8], [9], [10]). In this context, several projects have composed music while taking into account the limitations of electric hearing and have organized concerts such as "Noise Carriers" (Glasgow, United Kingdom, in 2007), "C4CI Grand Finale" (Southampton, United Kingdom, in 2011–2012), "Interior Design" (Melbourne, Australia, in 2010 [4]), and "musIC 1.0" [5] (Barcelona, Spain, in 2011). The goal of these concerts was to understand the differences in both appreciation and perception between NH and CI users in live music concerts. Moreover, these concerts serve to increase awareness about hearing loss and the technologies available to rehabilitate hearing loss by means of HAs and CIs. Another goal was to motivate CI users to participate in music related social activities together with friends and relatives and to have the opportunity to share a musical experience with other CI users and normal-hearing listeners.

The "Interior Design" project consisted of several seminars where musicians, CI users, and engineers met to compose, in its majority, electroacoustic music. Electroacoustic music is a genre developed around the middle of the 20th century that incorporates electric sound production techniques into compositional practice. It moves on the limits of what is considered music because it confronts common sense approaches to music based on melody and harmony, which are actually the dimensions most severely impaired when listening through a CI. The perception of the compositions was evaluated through post performance questionnaires by normal-hearing listeners and CI users in live concerts. The questionnaires had sections to collect qualitative and quantitative data about technical, affective, and cognitive reactions, as well as demographic data from audience members. The results from the questionnaire of these concerts indicated that both normal-hearing listeners and CI users, in general, considered the events a success. The results revealed similar responses from both groups in terms of interest, enjoyment, and musicality, although melody and timbre perception were rated lower by CI users while their ratings of percussion pieces were typically higher [4].

The "musIC: music for cochlear implants" project consisted of understanding how music is perceived live in a group of CI users. In a first phase of the project, a group of CI users performed tasks on instrument identification (timbre), melodic contour identification and rhythm perception. Additionally, each study participant filled questionnaires about their musical background and they provided us with collections of the music they usually listen to. In a second phase, a series of three seminars were organized where music composers, the same group of CI users and researchers interacted with each other to create music compositions. The information collected in the first phase was useful for the composers to better understand the limitations in

music perception with CIs as well as the large intrasubject variability. Finally, in a third phase, the compositions were interpreted in a concert where the appreciation of the music was evaluated by means of a questionnaire.

The current study reports the results obtained from the musIC 1.0 concert. The new compositions were presented during the musIC 1.0 concert which featured a combination of electroacoustic, pop and traditional music interpreted with acoustic instruments. The use of both acoustic and experimental electroacoustic music was chosen to explore the perceptual differences between NH listeners and CI users for a wide variety of acoustic inputs. The choice of different music styles during the concert was important to evaluate the responses of the different audience groups without being influenced by music genre preferences and to track the responses under very different acoustic scenarios. Some compositions used synthetic sounds with a clear fundamental frequency to facilitate the perception of melodies; or used simple passages at the beginning of the piece that were repeated with increased complexity trying to guide the listeners through the composition; or were mostly based on rhythmical structures and relied less on melody while using single instruments to reduce its complexity.

2 Methods

This section presents the music pieces composed for the musIC 1.0 project and the questionnaire used to evaluate them. The concert was organized in Barcelona (Caixa Forum) on February 9th, 2013.

2.1 Music Pieces

Piece 1 – Levit 1, composed and interpreted by Luis Afinador (Instrumentation: Piano). This piece covers different dimensions (rhythm, timbre, frequency, envelopes, harmonic complexity, clearness, etc.) with the goal to be perceived as pleasant as possible while generating musical challenges to the audience.

Piece 2 – Levit 2, composed and interpreted by Lusi Afinador (Instrumentation: Piano). Melodies based on relatively low fundamental frequencies were created, in the range from 500 to 1000 Hz, giving CI users the possibility to perceive the fundamental frequency through both temporal and place pitch. The composition uses a high variation in all dimensions while keeping a continuous path through the piece.

Piece 3 – Coclear, composed by Alejandro Civilotti (Instrumentation: Violin, viola, cello and flute). The piece approaches music to CI users based on research and creation. The composer of this piece is a CI user himself and had the idea to use music as a means to explain his own hearing loss experience followed by his CI implantation and rehabilitation.

Piece 4 – Tambora composed by Alejandro Fränkel (Instrumentation: Violin, viola, cello, flute and electronic). This piece is based on rhythmic structures. The idea of the composition is to use all instruments as a huge drum inside a folkloric ritual. The piece is therefore based on rhythm structure, the dimension of music that CI users can perceive best.

Piece 5 – El discurso vacío composed by Alejandro Fränkel (Instrumentation: Singing voice and electronic). Singing voice through vowels and consonants with large intervals, only the unusual things are important for the soul. The reality is something far away that approaches with infinite slowness should you be patient.

Piece 6 – Almost New Places, composed and interpreted by Sergio Naddei (Instrumentation: guitar). One of the most important aspects of hearing is balance and orientation in space. Hearing is therefore like visiting a space in which shapes and features can be appreciated by means of the air flow inside. Losing hearing and receiving a CI to hear again is equivalent to revisit spaces and places known in the past. After CI implantation, however, everything is perceived in a different way, sensations are overlapped with memories and the places area new again, or almost new.

Piece 7 – Almost New Spaces, composed and interpreted by Sergio Naddei. (Instrumentation: Reactable). Similar concept as "Almost new Places", this time interpreted using the ReacTable. A new synthesized sound was created in which the fundamental frequency is first presented as a pure tone and the harmonics appear sequentially over time creating notes that start simple (pure tone) and increase in complexity over time.

Videos of the recorded compositions can be found in [5]. Some music pieces were accompanied with real time visualizations trying to provide with additional information about pitch and rhythm to the audience.

2.2 Music Perception and Appreciation Questionnaire

The music questionnaire was adapted from the one designed by Innes-Brown, et. al. (2012) [7], that assessed cognitive aspects of music (interest, emotion and understanding), enjoyment, and technical aspects (melody, rhythm and timbre perception) for each of the new compositions. Each question was answered by means of a Likert scale ranging from 1 (very unsatisfied) to 5 (very satisfied). Each music piece used different instruments and technologies which impacted the ratings in different ways.

The questionnaire consists of two parts, the first one includes questions to quantify the musical experience and hearing performance of the listener. The second part consists of eleven questions about technical aspects, cognitive responses and engagement with the music for each composition. Finally, the questionnaire includes an open question where the participants can give additional information about the composition. The questionnaire was filled during the concert after each composition. The questionnaire consisted of 12 questions with an additional question:

1. The piece was very interesting
2. The piece was boring
3. I did not understand the piece
4. The piece was very enjoyable
5. I found the piece very musical
6. I could perceive well the melody
7. I could perceive well the rhythm
8. I can distinguish the instruments
9. The visualization was helpful to understand the melody
10. The piece made me feel calm
11. The piece made me feel happy
12. Did you use the magnetic loop?

Open question: Tell me your opinion about the current music piece.
Answers were based on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

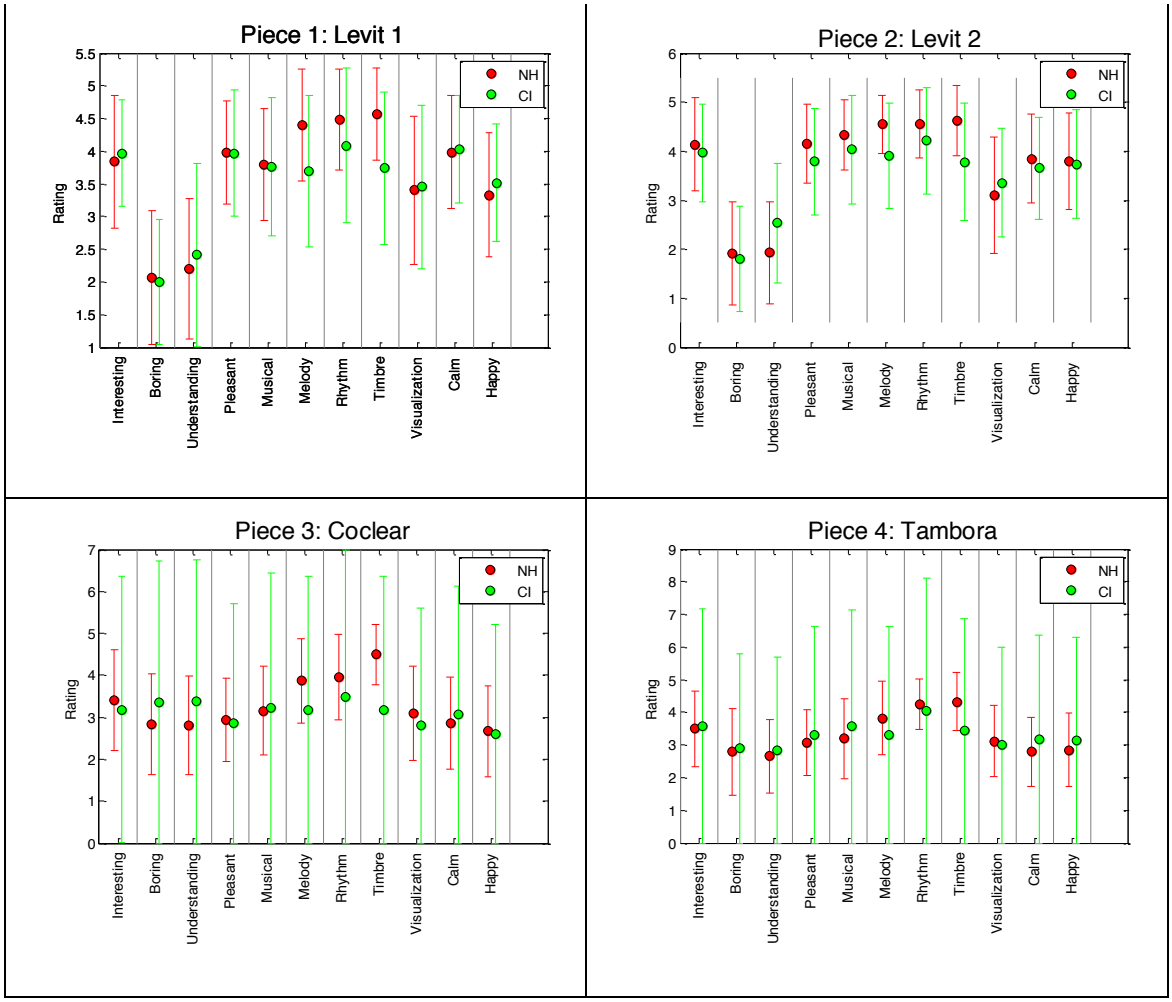
3 Results

253 people attended the concert musIC 1.0 and 133 filled the music appreciation questionnaire (86 NH, 37 CI users and 8 hearing aid users). Due to the low number of hearing aid users participating in the concert, this group was excluded from the analysis. From the 86 NH listeners, 37 were selected to match the age, musical experience and gender of the CI group.

Table 1. Summary demographic data for subsample of NH listeners and CI users for musIC1.0. Music ability was derived from part 1 of the questionnaire and ranges between 1 and 5.

musIC 1.0	**NH Group**	**CI Group**
N (females)	37(24)	37(16)
Median age in years	54	46
Music ability	1.47	1.27
Unilateral CI	NA	28
Bilateral CI	NA	9

Figure 2 presents the mean ratings of the questionnaire for each piece for the CI user group (in green) and for the NH group (in red).


Piece 1: Levit 1
Piece 2: Levit 2
Piece 3: Coclear
Piece 4: Tambora
Rating
NH
CI
Interesting
Boring
Understanding
Pleasant
Musical
Melody
Rhythm
Timbre
Visualization
Calm
Happy

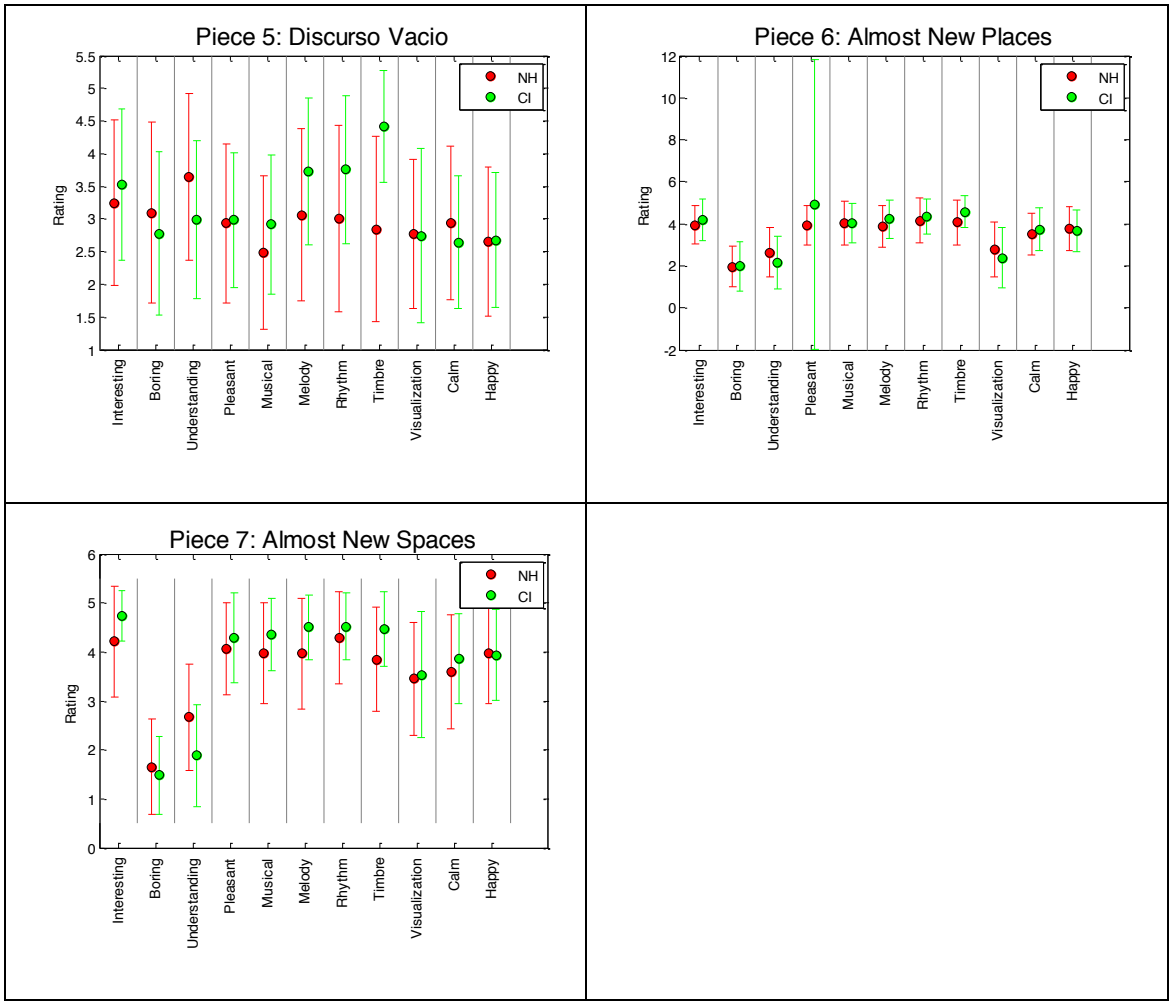
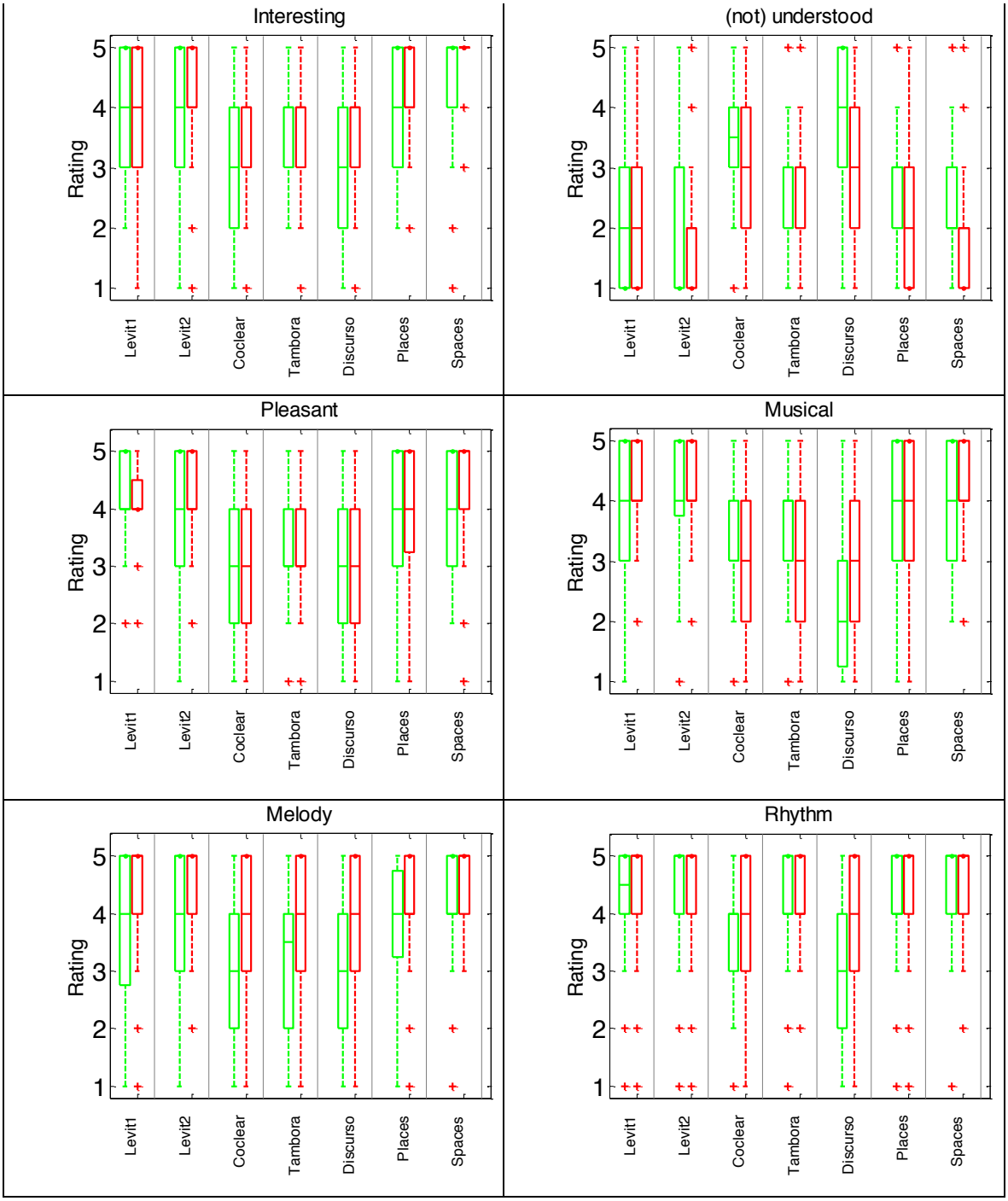


Fig. 1. Mean ratings from the 11 survey items for each composition of the musIC 1.0 project.

Figure 3 presents the results of the questionnaire for each of the 8 questionnaire items across music compositions.


Interesting
(not) understood
Pleasant
Musical
Melody
Rhythm
Rating
1
2
3
4
5
Levit1
Levit2
Coclear
Tambora
Discurso
Places
Spaces

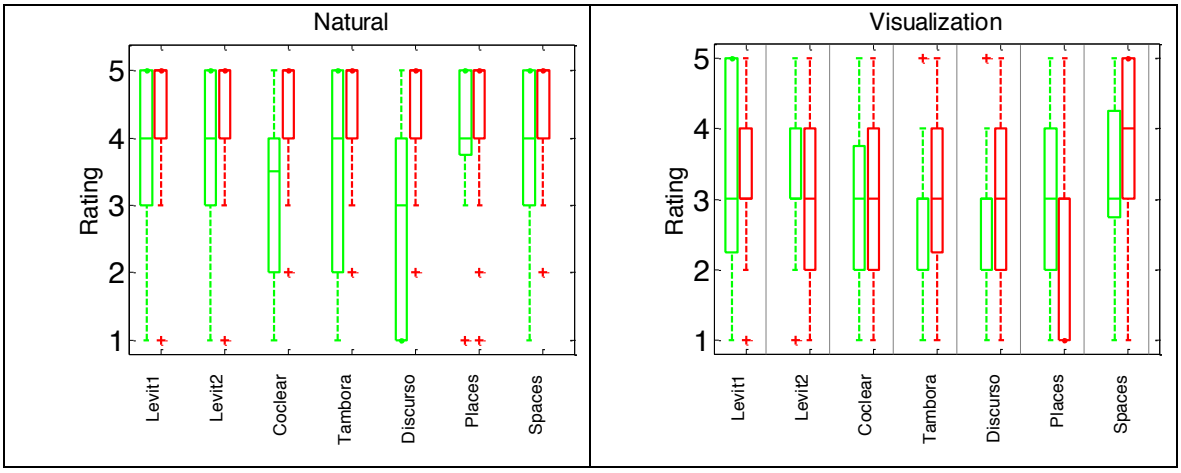


Fig. 2. Distribution of responses for the 8 questionnaire items analyzed for musIC 1.0 and for each composition.

3.1 Statistical analysis of selected items of musIC 1.0

We followed the same statistical procedure as described by Innes-Brown, et. al. (2012) [7]. From the eleven questionnaire items, only eight were selected to explore the three main areas of interest for further analysis.

The overall effect of group was investigated for each of the eight items using the Mann-Whitney test. The Friedman ANOVA was used to assess the effect of each piece and the Wilcoxon test was used to investigate main effects. Bonferroni correction was applied to compensate for multiple comparisons leading to a significance value of $p < .006$.

A) Cognitive response to the music

Item 1 Interest: 'The piece was very interesting': There was no significant effect of group (p=.039). However, the main effect of piece was found to be significant, X^2=52.349 and p = .000 for NH listeners, and X^2=30.76 and p = .000 for CI users. Post hoc tests revealed that the ratings for interest were significantly higher for 'Levit1', and 'Almost new spaces' for CI and only 'Almost new spaces' for NH.

Item 3 Not-Understanding: 'I did not understand the piece': There was no significant effect of group (p = .010). However, the main effect of piece was found to be significant; X^2=34.10 and p = .000 for NH, and X^2=30 and p = .000 for CI. Post hoc tests revealed that the ratings for non-understanding were significantly higher for 'El discurso vacío', and 'Coclear' for CI users, and 'Tambora' and 'Coclear' for NH listeners.

B) Engagement with the music

Item 4 Pleasure: 'The piece was very enjoyable': There was no significant effect of group (p =.359). Significant effects were found for piece, X^2=69.84 and p = .000 for NH and X^2=49.45 and p = .000 for CI. Post hoc tests revealed that the ratings for enjoyment were significantly higher for 'Levit1', 'Levit2', 'Almost new places' and 'Almost new spaces' for both NH and CI.

Item 5 Musicality: 'I found the piece very musical': There was no significant effect of group (p =.676). Significant effects were found for piece, X^2=50.18 and p = .000 for NH and X^2=45.34 and p = .000 for CI. Post hoc tests revealed that the ratings for musicality were significantly higher for 'Levit2', 'Almost new spaces' and 'Almost new places' in descending order for NH, and 'Levit2' and 'Almost new places' for CI.

C) Technical aspects of the music

Item 6 Melody: 'I could perceive well the melody': The main effect of group was found to be significant, U= 290.5 and p = .000. Significant effects were found for piece, X^2=33.22 and p = .000 for NH and X^2=50.18 and p = .000 for CI. Post hoc tests revealed that the ratings for melody were significantly higher for 'Levit1' and 'Levit2' for NH, and 'Levit2', 'Almost new places' and 'Almost new spaces' for CI.

Item 7 Rhythm:' I could perceive well the rhythm: There was no significant effect of group (p=.031). Significant effects were found for piece, X^2=29.07 and p = .000 for NH and X^2=32.49 and p = .000 for CI. Post hoc test revealed that the ratings for rhythm were significantly higher for 'Levit1' and 'Levit2' for NH, and 'Levit2', 'Tambora', 'Almost new places' and 'Almost new spaces' for CI.

Item 8 Timbre recognition: 'I can distinguish the instruments': The main effect of group was found to be significant, U= 199 and p = .000. Significant effects were found for piece, X^2=32.49 and p = .000 for NH but not for CI (X^2=3.193, p = .784). Post hoc tests revealed that the ratings for timbre recognition were significantly higher for 'Almost new places' and 'Almost new spaces' for CI users. No significant effects could be observed for NH listeners.

Item 9 Visualization: There was no significant effect of group (p = .377). Significant effects were found for piece, X^2=20.02 and p = .003 for NH but not for CI users (X^2=14.52 and p = .024). Post hoc tests revealed that the ratings for visualization were significantly higher for 'Levit2', 'Coclear', 'Almost new spaces' for NH, but no significant differences could be observed in CI users.

In summary the results show no significant difference in the ratings given by NH listeners and CI users on measures related to enjoyment such as interest or understanding for most of the pieces. No significant differences for measures related

to emotions such as calm or happiness were found. However, significant differences in the ratings given by CI users and NH listeners were observed for the questions related to the technical aspects of the music. CI users rated melody, timbre and rhythm perception significantly lower than normal hearing (NH) listeners. These results are in agreement with previous results using similar methods by Innes-Brown, et. al. (2012) [7].

4 Discussion

This work presented a subjective evaluation of music pieces in a live concert extending the work of [14] with new compositions. The results from the questionnaire indicated that CI users and NH listeners obtained similar enjoyment for the different compositions. As expected, the group of CI users rated melody, timbre and rhythm perception significantly lower than the group of NH listeners, however no significant group differences were found on measures of enjoyment and cognitive aspects such as interest, emotion and understanding. In general, CI users preferred simple music with clear melodies and rhythm complexity. In this context, technologies that try to emphasize the main melody of music piece or that try to simplify music are promising to make music more accessible for CI users (see [8], [9] and [10]).

The visual contribution to appreciation was not considered in this work. Future research should be able to disentangle acoustic from visual effects. One possibility would be to reproduce these concerts virtually and provide audio-visual or acoustic-only information.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2177/1—Project ID 390895286; musIC 1.0 received sponsoring from Advanced Bionics. The author would like to thank Dr. Hamish Innes-Brown for his advice and support and all subjects who participated in the musIC 1.0 concert and seminars.

References

1. H. J. McDermott, "Music perception with CIs: A review," Trends Amplif., vol. 8, no. 2, pp. 49–82, 2004.

2. V. Looi, K. Gfeller, and V. Driscoll, "Music appreciation and training for cochlear implant recipients: A review," Semin. Hear., vol. 33, no. 4, pp. 307–334, 2012.

3. C. J. Limb and A. T. Roy, "Technological, biological, and acoustical constraints to music perception in cochlear implant users," Hear. Res., vol. 308, pp. 13–26, Feb. 2014.

4. A. Au, J. Marozeau, H. Innes-brown, E. Schubert, and C. J. Stevens, "Music for the CI: Audience response to six commissioned compositions," Semin. Hear., vol. 1, no. 212, pp. 335–345, 2012.

5. Music for Cochlear Implants. (2017). [Online]. Available: www.music4ci.com.

6. W. Buyens, B. van Dijk, M. Moonen, and J. Wouters, "Music mixing preferences of CI recipients: A pilot study," Int. J. Audiol., vol. 53, no. 5, pp. 294–301, 2014.

7. H. Innes-Brown, A. Au, C. Stevens, E. Schubert, J. Marozeau , New music for the Bionic Ear: An assessment of the enjoyment of six new works composed for cochlear implant recipients , Proceedings of the International Conference on Music Perception and Cognition, July 23-28, 2012, Thessaloniki, Greece.

8. W. Nogueira, A. Nagathil and R. Martin (2019), Making Music More Accessible for Cochlear Implant Listeners: Recent Developments, in IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 115-127.

9. T. Gajecki and W. Nogueira (2018), Deep Learning Models to Remix Music for Cochlear Implant Users, Journal of the Acoustical Society of America, 143, 3602.

10. A. Nagathil, C. Weihs, and R. Martin, "Spectral complexity reduction of music signals for mitigating effects of cochlear hearing loss," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 3, pp. 445–458, 2016.

Embodied Cognition in Performers of Large Acoustic Instruments as a Method of Designing New Large Digital Musical Instruments

Lia Mice and Andrew P. McPherson *

Centre for Digital Music, Queen Mary University of London,
Mile End Road, United Kingdom E1 4NS
l.mice@qmul.ac.uk, a.mcpherson@qmul.ac.uk

Abstract. We present The Large Instrument Performers Study, an interview-based exploration into how large scale acoustic instrument performers navigate the instrument's size-related aesthetic features during the performance. Through the conceptual frameworks of embodied music cognition and affordance theory, we discuss how the themes that emerged in the interview data reveal the ways size-related aesthetic features of large acoustic instruments influence the instrument performer's choices; how large scale acoustic instruments feature microscopic nuanced performance options; and how despite the preconception of large scale acoustic instruments being scaled up versions of the smaller instrument with the addition of a lower fundamental tone, the instruments offer different sonic and performative features to their smaller counterparts and require precise gestural control that is certainly not scaled up. This is followed by a discussion of how the study findings could influence design features in new large scale digital musical instruments to result in more nuanced control and timbrally rich instruments, and better understanding of how interfaces and instruments influence performers' choices and as a result music repertoire and performance.

Keywords: Embodied Cognition, Digital Musical Instruments.

1 Introduction and Background

When interacting with an interface not only does the performer move their body to control the interface, the interface design and affordances control the way the performer moves their body. This paper introduces The Large Instrument Performers Study, an interview-based study with performers of various large acoustic instruments, and discusses the results, analysed through the thematic analysis methodology. The study findings are analysed in terms of embodied music cognition, affordances and idiomatic writing to show ways that size-related aesthetic features of large acoustic instruments shape the performer's choices while improvising, composing and performing repertoire.

* Research supported by EPSRC under grants EP/L01632X/1 (Centre for Doctoral Training in Media and Arts Technology) and EP/N005112/1 (Design for Virtuosity).

Through elucidating the ways in which the size of large instruments influence performance, instrument designers can learn about the ways large instruments are more than small instruments scaled up, and consider the impact of size-related affordances when designing new instruments. Now that DMIs can be any shape and dimension, and performed with virtually any gesture, exploring the impact of instrument scale on performers choices is useful for Digital Musical Instrument (DMI) designers when deciding what size to create instruments. As music production increasingly takes place in the home (due to faster consumer computers capable of running professional grade digital audio workstations), there is a trend in commercial DMI design of scaling down instruments and interface dimensions, resulting in smaller and smaller 'desktop' instruments such as the Korg Volca series of miniature synthesizers. But what is lost when an instrument is scaled down? More research is needed to understand the true impact of an instrument's scale and dimensions on music creation and performance.

Leman's embodied music cognition theory provides an interesting framework for exploring the impact of an instrument's scale and dimensions on music creation and performance, arguing that our bodily interactions shape our perception of music [1]. In the context of musical instrument performance, the body's 'vehicle' for mediation is musical gestures, which *"have an important experiential component that is related to intentions, goals, and expression."*[2] Not only are musical gestures linked to musical intentions, they are also linked to cognitive processing of the sounds they create, and in this way physical interaction with instruments involving gesture/action consequences changes our performance gestures and choices, and therefore our thinking.

Expanding or contracting the physical dimensions of an instrument results in changes to the musical gestures. In the taxonomy of affordance theory [3] [4] [6], it could be said that the size of an instrument influences its affordances, that is the possibilities, such as the gestural language for performance. Additionally, as De Souza argues, affordances offer the performer 'distributed cognition' in that an instrument may 'know' things for the performer [5]. In this way, the performer does not need to know every detail about the instrument to play it. As Magnusson [7] illustrates, the piano knows a pitch class represented by each of its keys. By only offering the tones created by pressing the piano keys, and not all the microtones in between, the embedded knowledge contained in a piano forms a 'script' that influences compositions created on the instrument. It follows that 'distributed cognition', or as Magnusson [8] calls it 'material epistemology', not only offers affordances but also constraints, and it is therefore through both that instruments elicit influence on performer's choices.

Tuuri et al. [9] argue that an interface enforces 'experiential control' on a user through 'push' effects (affordances that result in the user feeling the technology guides or constrains their embodied interaction) and 'pull' effects (affordances that result in the user feeling they are in control of the technology). Jack et al. [10] provide evidence of 'push' and 'pull' effects of a DMI's design on musical gestural interaction, showing that performers optimise their gestures to correspond with the sensing modalities of the instrument.

It can therefore be argued that the gestural language for performing an instrument is governed by the 'push' and 'pull' effects of the instrument's affordances and constraints. De Souza explores the link between affordances and distributed cognition, proposing 'idiomatic music' as those compositions which feature *"characteristic patterns that cannot be predicted by grammatical rules alone"*, arguing these characteristic patterns are the result of players interacting with the affordances of the instrument, composing not on a note-by-note basis but also through selection of 'ready-made sequences' on offer [5, page 77]. This music that is *"suited, adapted, and optimised for an instrument"* is what Tanaka [11] refers to as 'idiomatic writing', and is therefore the result of the physical affordances of the instrument. Huron and Berec [12] show idiomatic writing for an instrument can become less idiomatic if the circumstances change, observing that trumpet players find it more difficult to perform trumpet repertoire that is shifted in key or tempo so as to alter key fingerings and duration of breath.

The size of the instrument changes its relationship to the body and therefore its affordances, and in turn influences the idiomatic music of the instrument. However, more research is required to fully understand the extent of this influence as well as other factors that may be at play. In particular, the preceding reference raise the questions of what circumstances DMI designers can control and change, and the resulting impact on DMI repertoire and performance.

2 The Study

The Large Instrument Performers Study was designed to explore the possible impact musical instrument scale and dimensions may have on the performance of composed and improvised repertoire on the instrument by identifying affordances specific to large acoustic instruments, and how these affordances impact the performer's choices. The study consisted of one-on-one interviews with seven instrument performers who are trained on physically large instruments (see Figure 1). Some participants were trained on more than one instrument of an instrument family in which one instrument is larger than the other, for example baritone saxophone and tenor saxophone. The interviews lasted up to one hour. During the interviews, questions about performance technique and repertoire were asked and participants were encouraged to perform their instrument(s) as examples arose. The interviews were videoed and took place either in a professional music studio, on campus at Queen Mary University, or over Skype.

The participants were asked questions designed to reveal how the performers respond to effects introduced by the large scale of the instrument, such as physical navigation challenges, the additional physical effort required to perform larger instruments, and the relative changes in tone, timbre, volume and intensity encountered when performing repertoire on a large instrument versus a smaller similar instrument.

Specific questions asked included: Which techniques/patterns require you to move the most? Which techniques/patterns require you to move the least, or require microscopic precision? How long can you perform the instrument before

you are too tired to continue? What causes the fatigue? How do you think the instrument influences the music you make when improvising? Would you improvise in the same way on a different instrument? What is an example of well written music for your instrument? How would it change if you performed it an octave higher or on another instrument?

The performers were also shown repertoire composed for cello, "Cello Suite no. 1 in G Major" (all movements) by J. S. Bach, and asked what issues they would encounter if they attempted to perform it on their instrument. The videos were manually transcribed and the transcription data analyzed following a thematic analysis methodology [13]. Codes emerged through an iterative process that took a theory-driven approach [14], in that the raw interview data was examined for trends and correlations that relate to the theories of embodied music cognition, affordances and idiomatic writing. Four iterations of coding were performed resulting in a codebook that was updated and refined at each coding iteration.

3 Results

3.1 Thematic Analysis Codebook and Overarching Themes

The codes that emerged from the thematic analysis methodology were organised by the grouping of codes that shared a theme. Figure 2 presents an overview of the codebook structure and which participants commented on each code.

At a high level, we noticed a differentiation in the themes between those that describe instrument characteristics, and those that illustrate performer reactions to those characteristics.

In the context of how size-related affordances impact performer choices, the codes reveal both trends and individual insights that illustrate how large acoustic instruments impose fatigue issues on the performer influencing their decision of how long or whether to perform the instrument at all; how timbral variations across registers influence choices performers make when improvising on the in-

Participant Number	Primary Large Instrument Played	Other Instruments Played	Primary Style
P1	Contrabassoon	Bassoon, double bass, electronics	Contemporary, ambient
P2	Contrabass clarinet	Clarinets (soprano, bass, alto), flute, guitar, piano, saxophones (soprano, alto, tenor, baritone)	Contemporary classical, experimental
P3	Organ	Piano, soprano clarinet, voice	Classical, renaissance
P4	Contrabass flute	Flutes (bass, alto, concert, piccolo), recorder, piano	Contemporary
P5	Gyil	Percussion, drum kit, piano, guitar	World jazz
P6	Tuba	Guitar, gong, self-designed mechanical instruments	Metal
P7	Baritone saxophone	Saxophones (alto, tenor, soprano)	Jazz

Fig. 1. Study Participants

strument; how micro-level control and design of large instruments can result in substantial changes to the sound, influencing new performance techniques.

The interview content relating to techniques and repertoire performed on large instruments was categorised under three themes: idiomatic, easy and natural; unidiomatic, difficult and unnatural; and virtuosic or impressive composition. Comparing the insights that fell into one or more of these themes resulted in interesting insights into the differentiation between what is easy, natural, idiomatic and/or virtuosic in the context of idiomatic writing for large instruments.

3.2 Influence of Size and Weight on Performance Fatigue

Six out of the seven interviewees identified the cause of performance fatigue to be uniquely related to the instrument size. Causes included the instrument weight, the posture required to play the instrument due to its size, and extreme use of diaphragm/core muscles to support the air column and air pressure required to perform large scale woodwind instruments. As a result, five out of seven of the participants use a device or performance method designed to minimise performance fatigue caused by the instrument's weight.

In some cases the instrument's size and/or weight influences whether the performer chooses to perform the instrument at all. P4 commented she often opts not to perform with the contrabass flute at improvisational concerts because carrying the contrabass flute limits her ability to travel with more than one flute, whereas if she selects a smaller flute such as alto flute she has the option to also carry another flute such as concert flute or piccolo, offering her greater options at the concert. P6 said he seldom performs tuba in concert due to environmental concerns related to need to transport such a large instrument by car.

3.3 Timbral Variation Across Registers in Large Wind Instruments

Beyond identifying the aforementioned obvious size-related affordances of large instruments, the study identified a less obvious influence of the size of large wind instruments on composed and improvised repertoire. Large acoustic wind instruments are often designed to have a rich tone in the lower register. This feature is a result of the instrument having a very large pipe/sound chamber. Activating the entire chamber will result in the lowest, most resonant tone. Playing in higher registers uses smaller sections of the chamber, resulting in more airy, frail tones in the higher registers. These unusual upper tones are more difficult to perform in tune because more air pressure is required (due to the instrument's size). Maintaining a steady pressure at the intensity required is a difficult task for even the most advanced players.

Although a byproduct of the instrument's design, the unusual tones in the upper registers can become an interesting aesthetic resource to draw on when composing and improvising on the instrument. The study results indicate that the unique tones of both the upper and lower registers influences performer choices through embodied cognition and 'push' effects.

Code	Mentioned By Participants						
	P1	P2	P3	P4	P5	P6	P7
Impact of Size, Weight or Fatigue of Large Instruments on Performers							
Which technique/passage makes the performer move the most	x	x	x		x	x	x
Fatigue	x	x	x	x	x		x
Weight		x		x		x	x
Strength required to perform instrument	x	x	x		x	x	
Size		x					
Timbral Variation Across Registers In Large Instruments							
Choosing difficult techniques for sonic gratification	x						
Effects of variation across register on repertoire arrangements			x	x		x	x
Effect of playing in different registers on idiomaticity	x	x		x		x	
Influence of timbral variation on repertoire	x	x	x	x			x
Instrument is designed to have a strong bottom register	x		x	x			
Micro Scale Within Macro Scale of Large Instruments							
Microscopic design that has a large effect	x						
Microscopic gestures that have a large effect	x			x		x	
Improvising or Composing on Large Instruments							
The feel of the instrument changes how I improvise		x					
What performer doesn't play when improvising		x		x			
What performer plays when improvising		x		x			
Idiomatic, Easy or Natural to Perform on Large Instruments							
Idiomatic techniques		x	x	x			
Peformance of idiomatic music	x	x			x	x	x
Composition relating to idiomaticity	x	x	x	x	x	x	x
What is easy to play on the instrument	x	x		x	x		
What makes music idiomatic for this instrument	x			x		x	x
Unidiomatic, Difficult or Unnatural to Perform on Large Instruments							
What is difficult to play	x					x	
What makes a composition unidiomatic	x	x		x		x	x
Examples of unidiomatic compositions		x	x	x		x	x
Performances of unidiomatic compositions		x		x		x	x
What is more difficult to play than it seems		x		x			
Virtuosic or Impressive Composions for Large Instruments							
Video of virtuosic composition	x					x	
Exampes of virtuosic or impressive writing	x	x	x	x			
What makes a composition virtuosic for this large instrument		x	x	x			
Performing a Different Instrument's Repertoire on Large Instrument							
Performing repertoire intended for a different instrument is possible		x	x	x	x		
Performing repertoire intended for a different instrument is not possible	x	x	x	x	x	x	

Fig. 2. Thematic Analysis Codebook

All performers interviewed improvise on their instrument. When asked what they often play when improvising, four out of seven interviewees mentioned drawing inspiration from the timbral variation across registers. Composing and performing improvisations that are influenced by this aesthetic is an example of embodied cognition, as the performers are making specific choices based on the instrument's affordance of different tone colours at each register.

P1 described that when performing tones in higher registers of the contrabassoon *"the notes begin to get weaker"*, creating a unstable timbral quality that he makes use of when composing. *"In an orchestral setting, unless you want to specifically exploit this change in timbre in taking the instrument up an octave it might be better to write (the same part) for a bassoon instead... In my own music however I swap octaves a lot specifically to introduce this slightly more frail sound."* By extending his compositions into the higher register for the purpose of utilising this 'frail' tone (rather than other compositional choices such wanting an ascending melody), P1 is revealing the 'push' effects of the timbral aesthetic of the high register. This is an example of embodied cognition in that P1's compositional choices are changed by interacting with the instrument.

P2 is also drawn to the timbral variation across registers on the largest version of the instrument he performs. He mentioned that the E-flat clarinet is designed to have a uniform tone across all registers. By contrast the contrabass clarinet is not, hence it affords more tonal options to the performer. He said the contrabass clarinet *"has a lot richer sounds and things that I can really do with it, whereas the clarinet has more of a certain kind of sound and it doesn't have the same richness and variation."* When asked to name a composition that feels natural to play on the contrabass clarinet, P2 nominated 'Dark Light' by Thanos Chrysakis [15] because it *"highlights the capabilities of the instrument."* Composed for contrabass clarinet, 'Dark Light' features long tones in both the low and high registers. P2 later mentioned that performing contrabass clarinet in the higher registers is more difficult and less precise than performing in the lower registers. *"The higher you go the more notes I have on a single fingering... so I can't move between them as quickly as I have to do it with my mouth rather than with my fingers, so the precision isn't the same."* We find it interesting that even though performance of contrabass clarinet is more difficult for tones in the higher register, P2 indicated the most natural composition to perform on the instrument (in Tanaka's terminology, an example of 'idiomatic writing' for contrabass clarinet) features many complicated tones in the higher registers.

Similarly, P4 said that performing the same part in different registers on the contrabass flute *"would probably make it more difficult. If it was going higher it would make it harder to play in tune."* P4 said this difficulty in performing the higher register in tune is a byproduct of the contrabass flute design, which was designed to optimise the lower register tone at the expense of the higher register tone. *"The smaller (flutes) are deliberately made to make them as even as possible. Whereas the bigger ones are deliberately made not to do that. Because for example if you're playing a bass flute and you're playing in a flute choir, what you want is a really strong bottom octave... (On the contrabass flute) you*

get much better resonance in the low register, but it's possibly a bit weaker and a bit out of tune in the higher register where you're not going to use it very much."

Additionally, P4 indicated that performing the same passage in different registers of the contrabass flute would result in changing the character of the music. P4 said *"the character between the octaves changes quite dramatically. They each have a very different tone colour... I think if you put it in a different octave it would definitely change the character of the music."* When asked what types of sounds and passages she performs when improvising on contrabass flute, her responses included *"slow melodic material, possibly in the different octaves."*

Notably, it is not only the weaker, higher register that influences the performer's choice to perform a tone despite its difficulty. P1 said *"What I love on the instrument (contrabassoon) is holding the low notes for a long time. But that is very difficult."* He explained that unlike performing a long bass tone on another instrument such as the piano which would require the relatively easy gesture of pressing a key with one finger, performing a long bass tone on the contrabassoon requires precise core control. *"The lower you get, the more control you need over a consistent flow of air."* Despite the effort, what P1 enjoys performing the most on the instrument is long sustained bass tones. He regularly features them in compositions, commenting *"if you're using that with something on top that is such a brilliant foundation."* When asked why he prefers to use the contrabassoon rather than for example an electronic instrument for sustaining long bass tones, P1 said *"The performative and aesthetic element is important to me. I like using big effort instruments to make relatively reduced music. I have been using smaller instruments for ease of travel and using pitch shifting pedals to take them down an octave and although the end result is almost the same sonically as playing on a bigger instrument, it changes the essence of the music."* That P1 prefers to perform such a difficult technique on the contrabassoon instead of an easier technique on a different instrument is another example of embodied music cognition as he believes that creating the (almost) identical tone on a different instrument *"changes the essence of the music"*, implying that the instrument, not the tone, is changing his perception of the music.

3.4 Microscopic Performance Techniques on Large Instruments

Three out of five of the performers of large scale brass and woodwind instruments commented that microscopic changes in the embouchure and air pressure can result in huge changes in the sound and tone quality.

On the contrabass flute, a millimetre change in air angle can result in large changes and even the sound being lost altogether. P4 said *"because the instrument is so big, the air has to travel, so even something very simple like changing octaves needs very precise control of the air stream. And the distance between the octaves feels much bigger than it would do on a smaller instrument. So if you're playing a normal flute it takes a lot less air, and also the notes feel much closer together because the tube length is so much smaller. So because of that, all of those intervals, everything gets expanded. So I think from that point of view*

you're using a lot of precision of the airflow all of the time... Literally, if the air goes one millimetre in one way you'll lose the sound or change the sound."

Similarly, changes to the contrabassoon reed can cause a large changes to the instrument's tone. P1 said *"If we want a soft reed we can sandpaper that down for ten seconds, that's going to get the instrument to behave in a completely different way from not really very much of a change. So yes even though it's very big, some of the small changes can have a profound effect on the instrument."*

As a contrabassoon part maker, P1 has discovered ways certain microscopic design changes can influence the overall character of the instrument. When creating his own crooks (also known as bocals, the thin s-shaped tapered tube that the reed connects to), P1 discovered microscopic changes to the angle of the taper result in each crook having a unique sound. When comparing one self-designed crook to another, P1 said that in one *"the inside gets bigger quicker than the other one. So this has the capability to play higher notes more reliably."* While the other crook may be less reliable in the upper tones, P1 added its own characteristic that can be desirable for certain repertoire. *"The trade off is this one has more fundamental in the low notes."* By refining his crook-making process he can now design characteristics into the tone of the instrument. *"If I know I want a darker sound I know what to change to make that."*

P6 described a microscopic tuba technique he uses when playing in unison with others to create a beating sound. *"Other players will play a solid note and then I'll slightly bend the pitch of my note to create beats and that's done by a minor change in the lipping. It's really subtle. It's probably a bit to do with the air pressure as well but it's mostly a small deviation in the lip."* The result is a perceived effect of the tone rhythmically starting and stopping even though each performer is playing one long tone.

3.5 Influence of Difficulty and Virtuosity on Idiomatic Writing

Interview data relating to improvisation, repertoire, gestural performance techniques and performing repertoire intended for different instruments revealed interesting insights into what makes a technique, pattern or composition more or less difficult, idiomatic, virtuosic or impressive to perform on large instruments. In many cases the results offered insights that contradict common preconceptions of idiomatic writing, such as the assumptions that idiomaticity is synonymous with ease of performance, and virtuosity is synonymous with difficulty of performance. The compositions P2 and P5 regard as the most idiomatic and/or natural to perform on their instruments also contain performance techniques they consider the most difficult. P5 said the most idiomatic music for the Gyil is the polymetric Degaari traditional music, elaborating *"holding both metres and being able to play between them - that's hard for me and I don't think that's virtuosic... And if people thought that it was hard when they're listening to me then I'm not doing it right."* We find it interesting that the factors that impress audiences about this music, such as its speed and use of the full range of the instrument, are not what make it difficult to perform, and the mental challenge posed by the polymetric groove is not necessarily a factor that makes it virtuosic.

Study Results	DMI Design Choices Influenced by Study Results
Impact of Size, Weight or Fatigue on Performers	
Weight	Consider ways DMI design may decouple size from weight, such as light-weight materials, a design that packs down into travel cases
Strength required to perform instrument	Consider how a DMI that requires physical strength to perform influences the material epistemological scripts of the instrument, and what is idiomatic to perform on it, for example resulting in slower tempos, recurring clusters of tones located near one another
Size	Consider ways to avoid size-related constraints of large DMIs such as open spaces/layouts performer can see past, translucent materials
Timbral Variation Across Registers	
Choosing difficult techniques for sonic gratification	Consider designing in 'easter egg' tones accessible via most difficult to perform gestures at the very limit of what is performable.
Effects of variation across register on repertoire or arrangements	Consider assigning one register as the most resonant and weaker tones in other registers; Offer access to many registers at once via many tones or choice of a scale with fewer tones per octave
Effect of playing in different registers on idiomaticity	Consider the ways that implementing sound design that varies across results in creating scripts of idiomatic music for the instrument.
Instrument is designed with a strong bottom register	Consider whether the DMI is intended for performance as a solo instrument or in ensembles
Micro Scale Within Macro Scale	
Microscopic gestures that have a large effect	Consider DMI designs that allow for microscopic gestures to result in a large sonic effect on the overall tone/performance of the DMI.
Improvising/Composing on Large DMIs	
Feel of the DMI changes improvisations	Consider how the strength and effort required to perform the DMI may influence the performances/compositions created on the DMI.
What is Idiomatic, Easy or Natural to Perform on Large Instruments	
Idiomatic techniques	Consider the impact of the 'push' and 'pull' effects of what tones or passages are created by the easiest to perform techniques.

Fig. 3. Suggested Guidelines for Implementing Findings into DMI Design

4 Discussion

Current ongoing trends in DMI performance research include effortfulness [16], physicality and whether controllerism/laptop music engages audiences [17]. We argue that while large DMIs engage more with the body and are more physical and visible than their smaller counterparts, more research is required to fully understand ways in which their size influences DMI music and performance.

Keeping in mind Magnusson's [7] notion of 'scripts' and 'material epistemologies', the hidden knowledge embedded in instruments that shape idiomatic writing, DMI designers can draw inspiration from the detail and variation of sonic features of acoustic instruments when creating sound design that inspires virtuosic composition on DMIs. The interviews with contrabass flute and contrabass clarinet performers show that the varying timbral qualities afforded by large wind instruments influence performers' choices when improvising on the instrument, as well as their decision of whether to perform the instrument at all

(in place of performing the smaller version with a more uniform tone across all registers). This indicates that non-uniformity of tone across registers is a strong aesthetic resource for compositional inspiration. DMI designers could consider implementing this characteristic not only large DMIs but DMIs of all sizes.

The observations from the Large Instrument Performers Study show that the 'push' effects of timbrally varied tones across registers influenced performers to make use of multiple registers while improvising and composing. That the participants chose to perform within the more difficult registers, even at the risk of discomfort or error, shows the extent to which performers value these tones. We argue these findings should encourage designers of DMIs of all sizes to consider the value of offering simultaneous access to multiple registers and varied sound design across registers, as well as microscopically precise gestural controls - even those initially unnatural or difficult to perform.

In light of the study findings that reveal large wind instruments respond to microscopic changes in gestural control and micro-scale design details, we argue that to reach new frontiers of virtuosic digital instrument performance and repertoire, large DMI designers should take into account the microscale within the macroscale. Scaling up the DMI to be larger is only the first step. Until large scaled DMIs match or exceed the nuanced precision of large acoustic instruments, large DMIs we will not reach their musical and performative potential.

Exploring human interaction with an instrument too large and complex to master was an approach taken by the group Sensorband (Atau Tanaka, Zbigniew Karkowski and Edwin van der Heide) with their architectural scale instrument SoundNet [11]. In the context of researching embodied cognition and idiomaticity, we argue there is more to be discovered from musical interactions with instruments designed to overwhelm the performer with its physicality. One physically overwhelming instrument discussed in The Large Instrument Performers Study was the contrabass flute, which requires so much breath support the performer can become dizzy. P4's expert insight into performing such a physically overwhelming instrument illuminated our discussion by providing a perspective from the extreme end of acoustic instrument performance.

4.1 Guidelines for Implementing Findings into DMI Design

Drawing on the findings on the study, Figure 3 outlines a series of design features that DMI designers could consider.

5 Conclusion and Future Work

We presented a study to examine the affordances of large acoustic instruments and their effect on performers choices. This study has shown us that large scale instruments are more than just small instruments scaled up; rather they are highly detailed, precise instruments that in many cases offer different sonic affordances than their smaller counterpart of the same instrument family. The findings revealed a series of interesting aesthetic design features of large acoustic

instruments, such as the timbral variation across registers and the microscopic precision of control, that have a strong influence on performers choices through embodied music cognition and 'push' effects. More research is required to understand the full impact of instrument size and scale on musical performance and composition, however this research offers initial insights to consider when designing new DMIs of all sizes.

References

1. Leman, M., Maes, P. J., Nijs, L., Van Dyck, E.: What is Embodied Music Cognition?. In: Bader, R. (ed.) Springer Handbook of Systematic Musicology, pp. 747–760. Springer, Berlin (2018)
2. Leman, M.: Musical Gestures and Embodied Cognition. In: Dutoit, T., Todoroff, T., d'Alessandro, N. (eds.) Actes des Journees d'Informatique Musicale, 5-7 (2012)
3. Gibson, J. J.: The Ecological Approach To Visual Perception. Houghton Mifflin, Boston (1979)
4. Norman, D.A.: The Psychology of Everyday Things. Basic Books, New York (1988)
5. De Souza, J.: Music At Hand. Oxford University Press, New York (2017)
6. Turvey, M. T.: Affordances and Prospective Control: An Outline of the Ontology. Ecological Psychology 4(3), 173–187 (1992)
7. Magnusson, T.: Of Epistemic Tools: Musical Instruments as Cognitive Extensions. Organised Sound 14(2), 168-176 (2009)
8. Magnusson, T.: Designing Constraints: Composing and Performing with Digital Musical Systems. Computer Music Journal 34(4), 62–73 (2010)
9. Tuuri, K., Parviainen, J., Pirhonen, A.: Who Controls Who? Embodied Control Within Human-Technology Choreographies. Interacting With Computers 29(4), 494–511 (2017)
10. Jack, R. H., Stockman, T., McPherson, A. P.: Rich Gesture, Reduced Control: The Influence of Constrained Mappings on Performance Technique. In: Proc. MOCO, pp. 15:1–15:8. ACM, New York (2017)
11. Tanaka, A.: Musical Performance Practice on Sensor-Based Instruments. In: Wanderley, M., Battier, M. (eds.) Trends in Gestural Control of Music. pp. 389–405. IRCAM-Centre Pompidou, Paris (2000)
12. Huron, D., Berec, J.: Characterising Idoimatic Organisation in Music: A Theory and Case Study of Musical Affordances. In: Empirical Musicology Review 4(3), 103–22 (2009)
13. DeCuir-Gunby, J. T., Marshall, P. L., McCulloch, A. W.: Developing and Using a Codebook for the Analysis of Interview Data: An Example From a Professional Development Research Project. In: Field Methods 23(2), 136-155 (2011)
14. Ryan, G. W., Bernard, H. R.: Techniques to Identify Themes. In: Field Methods 15, 85–109 (2003)
15. Chrysakis, T.: Dark Light. London (2017)
16. Bennett, P., Ward, N., O'Modhrain, S., Rebelo, P.: DAMPER: A Platform For Effortful Interface Development. In: Proc. NIME, pp. 273–276 (2007)
17. Bin, S. M. A., Bryan-Kinns, N., McPherson, A. P.: Hands Where We Can See Them! Investigating the Impact of Gesture Size on Audience Perception. In: Proc. ICMC (2017)

An Ecosystemic Approach to Augmenting Sonic Meditation Practices

Rory Hoy and Doug Van Nort

DisPerSion Lab
York University
rorydavidhoy@gmail.com, vannort@yorku.ca

Abstract. This paper describes the design and creation of an interactive sound environment project, titled *dispersion.eLabOrate*. The system is defined by a ceiling array of microphones, audio input analysis, and synthesis that is directly driven by this analysis. Created to augment a Deep Listening performative environment, this project explores the role that interactive installations can fulfill within a structured listening context. Echoing, modulating, and extending what it hears, the system generates an environment in which its output is a product of ambient sound, feedback, and participant input. Relating to and building upon the ecosystemic model, we discuss the benefit of designing for participant incorporation within such a responsive listening environment.

Keywords: Interactive Audio, Sonic Ecosystem, Deep Listening

1 Introduction

In contrast to fixed-media works for concert, the generation of a sonic environment for an installation context invites participants to traverse a space, wherein their action has amplified potential to modulate generated sound through manipulation of devices, interfaces, and the ambience of the room itself. The systematic formation and implementation of these interactive works is dependent upon the role of participants within the space (or lack thereof). Techniques range from a linear system flow wherein participant action directly drives generated output to a sonic ecosystem approach, wherein feedback mechanisms establish autonomous and self sustaining sonic activity.

This paper will explore the formation of *dispersion.eLabOrate*, an interactive sound environment which began as an augmentation on the form of the "Tuning Meditation", a text piece found within the practice of *Deep Listening* [1]. This meditative and performative context informed the aesthetic and design considerations employed within the system's development, due to its need to function as a collaborative member of the piece, rather than distracting from the focused listening context in which it was deployed. The system was developed with the design metaphor of an "active listening room" in mind, reacting both to participants and its own generated audio. The relationships established between human, machine, and ambient environment led to exploration of the ecosystemic approach presented by Agostino Di Scipio [2]. Contending with boundaries put in place by the classical

ecosystemic approach, *dispersion.eLabOrate* presents a model in which the human and the machine can act together in the generation of an ecosystem such that the blending of agency is achieved through the system's self/ambient observing behavior and the participant's ability to be present in this observation. We will discuss the need to bridge between methodologies for interactive sound environments, presenting an approach that extends the capabilities of a sonic ecosystem dynamically through participant input, resulting in spatially distributed parameter changes. These localized changes can be thought of as generating diverse locations within the sonic ecosystem, with input conditions resulting in distinct perceptual effects for both participants and the ambient sensing of the system.

2 Related Works

2.1 Ecosystemic Framework

Undertaken in Di Scipio [2], the challenge of generating a sonic ecosystem is engaged by questioning the nature of interactivity, by exploring the limits of "where and when" this occurs. Di Scipio notes that the majority of interactive systems employ a linear communication flow in which a participant's action is the singular cause of output. Di Scipio then presents an alternate approach in which the principal aim is the creation of a dynamical system which can act upon and interfere with the external conditions that define its own internal state. This approach decentralizes the primal importance of human agency in the space (apart from ambient noise) and grants the ability of self-observation to the system. Di Scipio describes this ability as "a shift from creating wanted sounds via interactive means, towards creating wanted interactions having audible traces"; and it is through these traces that compelling sonification can occur. This ideation of an audio ecosystem culminates in Di Scipio's Audible Eco-Systemic Interface project (AESI) project. This machine/ambience interrelationship is paramount and understood to function as "interaction", rather than the typical human/machine relationship. AESI emits an initial sound that is captured by two or more microphones in the room. Relevant features are extracted from this capture, which are then used to drive audio signal processing parameters. Measurements on differences between microphone signals are used as additional control values, and the internal state of the AESI is set through functions defined by this ecosystemic concept. The four functions achieving this are compensation (active counterbalance of amplitude with the ambient environment), following (ramped value chasing given a delay time), redundancy (supporting a predominant sound feature), and concurrency (supporting a contrasting or competing predominant feature).

These defining ecosystemic characteristics of equilibrium and adjustment are explored by Haworth [3], who suggests the need to update the ecosystemic model to reflect current broader thoughts on ecosystems, de-emphasizing stability and highlighting imbalance and disorder. Haworth identifies two distinct models, stemming from Di Scipio and Simon Waters. Di Scipio's form is a cyclical closed system in which traditional control structures of linear systems in interactive audio works are dismantled in favor of a self-regulated ambient sensing. Meanwhile, Waters moves away from tendencies to instrumentalise technology, instead highlighting the

role of human attention upon relations formed between each of the components within a generated ecology. Waters posits, "The notion of Performance Ecosystem enfolds all three concepts (performer, instrument, environment) and allows room for the undecideabilities of the virtual domain" [4], depicting this interrelated nature of ecosystemic components as primary over their intersection with the "virtual domain". While an extended examination of these two positions is beyond the scope of this paper, for the purposes of this discussion it is suitable to work from this relatively high-level distinction between the two. In so doing, the modified understanding of the ecosystemic model posed by Haworth [3] and the situated performance ecology of Waters [4] are most applicable to the system design of *dispersion.eLabOrate*. Incorporating aspects of system self-observation while explicitly designing around participants' attentional dynamics, the generated sonic ecosystem deals with the blending of influence between the system and environmental actors.

2.2 Deep Listening

The practice of Deep Listening was developed by composer Pauline Oliveros in the 1970's and refined into the 2000's. It is described by Oliveros as "a practice that is intended to heighten and expand consciousness of sound in as many dimensions of awareness and attentional dynamics as humanly possible" [1]. With a focus on embodied listening to internal and environmental stimuli, the practice integrates somatic awareness and energy exercises, listening meditations, and sound-making exercises that build upon an initial set of text-based pieces Oliveros created known as Sonic Meditations, with the Tuning Meditation (TM) being one of the earliest and most widely-practiced. The Deep Listening community has grown through regular workshops to include thousands of past and current practitioners, and is kept alive through certified Deep Listening instructors, including the second author.

3 System Description

The project was created in the DisPerSion Lab at York University in Toronto, an interdisciplinary research-creation space outfitted with a multichannel audio system. For *dispersion.eLabOrate*, 12 channels mounted on floor stands were employed, with positions chosen in order to mitigate extraneous feedback, while facilitating intended feedback between the generated audio and the array of ceiling mounted microphones. The array of 3x3 omnidirectional microphones ensures participant input is evenly sensed throughout the space. The TM asks participants to inhale deeply, exhaling on a note/tone of their choice for one full breath. On the following exhalation, participants will then match a tone that another has made. Next, a new tone should be held that no one else has made. This alternation between matching others and offering new tones repeats until a natural end point is reached, as determined by the group listening dynamic. In this project we also allowed participants to choose between noise or tone at each cycle. As this was the primary context in which the project was intended, all major aesthetic considerations and testing revolved around ensuring the piece could be performed without distraction. The role of the system is to extend the potential for the piece, rather than overtake it as a singular focus.

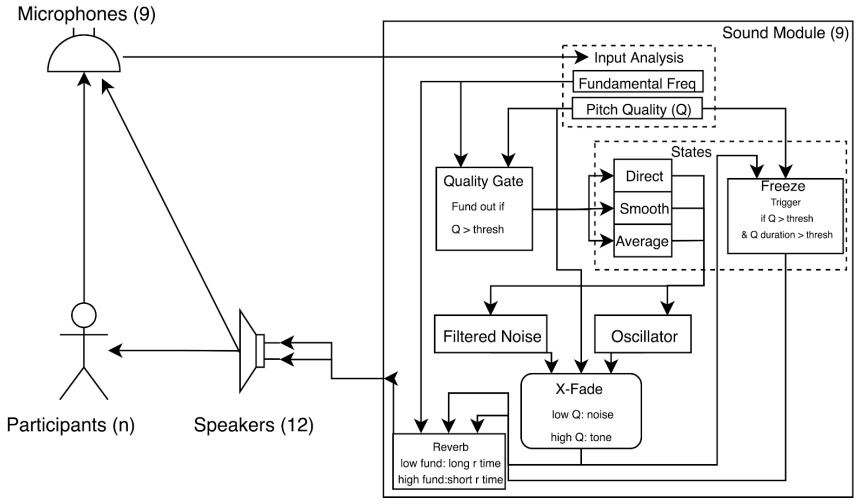


Fig. 1. System diagram of *dispersion.eLabOrate* depicting signal and data flow from microphones, through pitch analysis, to audio generation, and output to room

The audio is received by the computer via an audio interface connected to the microphones. Incoming audio is then accessed by Max/MSP, where the analysis and audio generation occurs. The system is comprised of 9 modules, one for each of the microphones in the array. Each module consists of a sinusoidal oscillator, as well as a white noise generator that is fed into a bandpass filter. The system's output is located spatially with regards to the location of the microphones within the room, placing each of the 9 output signals in relation to their input source. This placement promotes feedback at the localized level between a modules output and accompanying microphone, while additionally influencing adjacent output and microphone pairs. The modules contain states which alter the behavior of audio generation and its listening parameters. The four states are, *direct*, *smooth*, *average*, and *freeze*. These states differ in the way they map values to the module's oscillator and filter, and change parameters for data thresholding. States can be set individually for each module, allowing varied behavior within localized areas of the room. Each audio input is analyzed for fundamental frequency, and pitch quality (an estimation of analysis confidence). Fundamental frequency is calculated by the zsa.fund method [5] and pitch quality estimation is extracted using the yin algorithm [6]. Yin was not used for fundamental frequency tracking as it was found to increase feedback past a desired level, hence the use of the FFT-based method. The 9 separate modules receive the fundamental frequency and pitch quality from their respective microphone, which are then sent to the module's oscillator and filter. The fundamental is used as the desired frequency for the oscillator as well as the centre frequency for a resonant band pass filter. Values are only sent if a defined threshold for pitch quality is passed (default 0.2), and pairing this quality gate with a noise gate on the original microphone signal avoids having unintentional ambient stimulus/noise as input. Moving between ostensibly simple states results in a potentially drastic difference of behavior for the system's output. *Direct* causes the analyzed fundamental frequency

to be immediately reflected in the oscillator and filtered noise. *Smooth* sends values to the output sources ramped over time (default 50ms). *Average* sends out values to the sources after calculating a running mean during a given time window (default 200ms). *Freeze* implements spectral freeze and sustain techniques [7], triggering them when input passes a set pitch quality threshold and pitch quality duration (default 1s). In addition to gating data flow, the pitch quality value is used to crossfade between the two audio generation sources of each module. Low pitch quality is perceptually tied to "noisy" input stimulus, while high pitch quality will result from clear tones. When the quality value is low, output will be closer to the filtered noise. If the quality is high, output will be towards the generated pure tone of the oscillator. Thus the resulting output of a module is congruous with the timbral quality (ranging from tone to noise) at any given mic. Reverb was added to accentuate the spatial aspects of the audio generation and was also controlled by the analyzed fundamental frequency at the module level. Low frequency was mapped to a long reverb time, while high frequencies were mapped to a short reverb time.

4 Evaluation and Discussion

4.1 Tuning Meditation User Study

In order to systematically examine the perceived influence of *dispersion.eLabOrate* across its four states, a user study was conducted with five volunteers joining the two authors, for a total of seven participants. The TM ran five times in a row: first without *dispersion.eLabOrate*'s sensing to establish a "ground truth". The four following runs implemented the system states, moving through *direct, smooth, average,* and *freeze*. Participants were allotted time to write personal comments and rest in between each run. A survey was completed after the final run and before group discussion, to avoid biasing personal reflections on the experience. The survey utilized a five-point Likert scale, with the following questions for each run: Q1: During this piece/experiment, could you differentiate any electronic sound output from that of human performers? Q2: During this piece/experiment, could you recognize any tones/noises being matched (either yours or another person's) by another human performer? Q3: During this piece/experiment, could you recognize any tones/noises being matched (either yours or another person's) by electronic sound output? Q4: How confident are you about your recollection of run #N and related ability to answer these questions?

The responses show a trend in participants reporting less ability to recognize tones being matched by fellow humans (Q2) in successive runs. This may be a product of becoming more comfortable with the system as another participant/agent within the piece. This is supported by participant comments: noting in run #3 that the "electronics faded in background - less interested in triggering the electronics than using it as a source for unique tones" and in run #5 that "the electronics lost novelty (and) acted more as (a) participant in my mind". The same participant noted of run #3 that "the electronics held (the) same importance as other performers", whereas they earlier reported that they "spent (the) first few breaths figuring out what tones would trigger the electronics". Another participant noted in run #3 that "the machine felt like it was part of the sound field, but in a different way to the rest of the participants"

whereas by run #5 they noted that the run "had a very satisfying ending when the machine faded out with the group", pointing to its collaborative place within the piece. While the surveys required a recollection of every run from the 1.5 hour session, each participant reported high confidence in this recollection across every run. The general trend of recognizing less human matching (from the quantitative data) and increasing regard for the interactive system as an agent to be listened and responded to (qualitative comments) is quite interesting. This certainly must be related to an increasing familiarization with the system, but it also may be related to the specific ordering of the states: while all system output was normalized to the same volume (balanced to blend with group sound), state changes from runs 2-5 correlated with an increased sustain of system output due to state behavior. This greater sustain, and related self-regulatory feedback, seemingly contributed to the increased sense of presence reported, with participants noting that the sound was "less chaotic" and contributed to the larger environmental context of the experience. This speaks to the influence of the ecosystemic design on perceived agency.

4.2 Discussion

The "Tuning Meditation" Deep Listening piece is itself an emergent dynamical process that could be seen as an acoustic form of an interactive sonic ecosystem. When intersected with *dispersion.eLabOrate*, the result is a piece positioned within the ecosystemic model through shared human/technological influence. Due to the flexible number of participants that may take part in a performance/session of the TM, variances in voice density may be quite apparent or perceptually unnoticeable due to aligning breath cycles. "Feedback" is inherently present through the act of matching another's output and the ambient qualities of the piece are established by all participants acting to form a self-regulating system. Additionally the piece is ran until collective stimuli concludes, further positioning the importance of ambient content to drive the output of the human "system" established between participants, noted in the user study through the comments of the "machine" ending the piece in run #5. All of these participant interrelationships are extended through the addition of the generated audio of *dispersion.eLabOrate,* as behaviors not typically found in the original piece and "vocalizations" not achievable due to human physical constraints emerge from the system. This was evident in the user study through comments that regarded the environment as another agent, and has been further apparent to the authors across test sessions. Incorporating behaviors such as *freeze*, the system is able to sustain tones across gaps in participant stimuli, allowing continuous output to take place in the piece even within small groups. This was shown to have a noticeable positive effect on group coherence, with participants noting that the "interactive sound became more meaningful". While these extensions of human ability are present within the system, an important design consideration was that output was still bound to the activity provided by participants. System output is reliant on a "communal breath", as the cyclical deep exhalations on unique or matched tones drives the system's audio input. The system is at once an actor taking part in the meditation along with the other participants, as well as the generator of the environment in which it resides. Each of the system's states presents a different possible form that sonic ecosystems can take

Fig. 2. *dispersion.eLabOrate* was developed in the context of a project that explored different input sensing, media output displays and interaction designs for augmenting sonic meditations.

within an interactive audio environment. Where the *direct* state results in the real time modulation of input audio mapped to output found in systems that employ a linear communication flow, *smooth*, *average*, and *freeze* move the system's behavior away from this one-to-one mapping. *Smooth* results in a behavior that is clearly linked to, but perceptually disjointed from participant input. This state results in "audible traces", where generated output hangs in the environment and is perceivable over a duration of time. These dynamic gestures of sound lack stable forms and fluctuate around the system's input (to varying degrees given a certain delay time). Audible traces continue within both the *average* and *freeze* states. *Average* behaves similarly to *smooth* as its calculation window begins, and upon receiving a number of samples will begin to reach a steady-state and settle around a small range of tones. At the end of the averaging window, the system's output may jump drastically to the current input fundamental. This cycle of progressively static and eventually collapsing forms is again self-referential in relation to feedback detected by the microphones, modulated and informed by the input of participants within the space. *Freeze* became arguably the most consistently intriguing of the states, as hanging tones and rhythmic sustained patterns were formed as a product of a surpassed pitch quality threshold, in combination with surpassed specified quality duration. The frozen tones were also spatialized to the location of the microphones detecting them, placing the live system output and spectral capture of sound within the same point of emanation. Generated output possibilities including beating waves and cyclical "following" behaviors caused by new frozen tones being generated from past output, given their proximity to adjacent microphones and source positions.

Reverb acted in facilitating positive feedback within *dispersion.eLabOrate*, allowing the system to further obtain the self-observing behavior that is characteristic of sonic ecosystems. Reverberation time is tied to the incoming analyzed frequency of each of the microphones, where low frequency content results in a high reverberation time and high frequencies cause a very short reverberation time. If a continuous low tone were to be captured by the system, the reverb time would be quite large (~10 seconds). This continuous tone could then be disturbed by input at a higher frequency than previously generated, causing the output of the system to spike in frequency, reducing the reverberation time, and collapsing the generated sonic structure. This behavior reflects Haworth's perspective on sonic ecosystems, "which de-emphasizes stability and regulation in favour of imbalance, change and disorder" [3].

5 Conclusion and Future Work

Created as a system to augment the sonic output of the Deep Listening "Tuning Meditation", *dispersion.eLabOrate* drew upon an ecosystemic design approach in its methodology, aesthetic output, and system considerations. Approaching perceived sonic agency as a symbiotic relationship between human and machine output, the work succeeds in placing human actors as integral to and active in the analyzed room ambience. This active participation within the environmental ambience is reliant on the generated output from the system, informed by chosen states for varied or uniform system response. Through the states *direct, smooth, average,* and *freeze, dispersion.eLabOrate* sculpts the environment participants are engaged within, while becoming an active participant itself within the framing of the piece. Cycling through these module states illustrates the potential for multiple interaction paradigms and system outputs from simple mapping changes within a single environment, highlighting the complex role of collective human action in the presence of feedback as found within the ecosystemic approach. Currently the system has the capability to define localized behavior within the sonic ecosystem through its individual modules which are related to each of the microphones in the space. The dry/wet content of reverb was not connected to any input analysis feature for this project, yet incorporating a reactive nature to this parameter could yield perceptually interesting variations for dynamically defining the shape of the sonic ecosystem at a localized level. This could also be applied to the function and assignment of states at the module level, defining multiple sonic locations in which output and systemic behavior varies, yet their collective output and proximity coalesce into a cohesive sonic ecosystem. This could allow autonomous reactive changes to occur as a result of decision making from the system, as determined by the structure of an exercise such as a sonic meditation, or through participant input. Such dynamic localized behaviour (either pre-set conditions or reactive) points towards exciting applications of sculpted, diverse, and mutating sonic ecosystems for use through augmenting participatory listening/sounding pieces such as those found within the Deep Listening tradition.

References

1. Oliveros, P.: Deep Listening: A Composer's Sound Practice. iUniverse, Lincoln. (2005)
2. Di Scipio, A.: 'Sound is the interface': from interactive to ecosystemic signal processing. Organised Sound, vol. 8(3), pp. 269--277, United Kingdom (2003)
3. Haworth, C.: Ecosystem or Technical System? Technologically-Mediated Performance and the Music of The Hub. Electroacoustic Music Studies Network (2014)
4. Waters, S.: Performance Ecosystems: Ecological approaches to musical interaction. Electroacoustic Music Studies Network (2007)
5. Malt, M., Jourdan, E.: Zsa.Descriptors: a library for real-time descriptors analysis. Sound and Music Computing Conference (2008)
6. de Cheveigné, A., Kawahara, H. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, vol. 111(4), 1917--1930, (2002)
7. Charles, J.-F.: A Tutorial on Spectral Sound Processing Using Max/MSP and Jitter. Computer Music Journal, vol. 32(3), pp. 87--102, (2008)

Gesture-Timbre Space: Multidimensional Feature Mapping Using Machine Learning & Concatenative Synthesis

Michael Zbyszyński, Balandino Di Donato, and Atau Tanaka *

Embodied Audiovisual Interaction Group
Goldsmiths, University of London
New Cross, London, SE14 6NW, UK
m.zbyszynski@gold.ac.uk, b.didonato@gold.ac.uk, a.tanaka@gold.ac.uk

Abstract. This paper presents a method for mapping embodied gesture, acquired with electromyography and motion sensing, to a corpus of small sound units, organised by derived timbral features using concatenative synthesis. Gestures and sounds can be associated directly using individual units and static poses, or by using a sound tracing method that leverages our intuitive associations between sound and embodied movement. We propose a method for augmenting corporal density to enable expressive variation on the original gesture-timbre space.

1 Introduction

Corpus-based concatenative synthesis (CBCS) is a compelling means to create new sonic timbres based on navigating a timbral feature space. In its use of atomic source units that are analysed, CBCS is an extension of granular synthesis that harnesses the power of music information retrieval and the timbral descriptors it generates. The actual sound to be played is specified by a target and features associated with that target. In speech synthesis, the target is text. In audio resynthesis and "mosaicing" applications, the target can be another sound. In digital musical instrument (DMI) performance, the target may be sensor data or some representation of performer action, or gesture. The target may be of the same or different modality than the corpus, and it may have the same or different feature dimensionality.

CBCS performance systems until now have, on the whole, been implemented using dimensionality reduction. A subset of corporal features are projected into a low dimension space, typically Cartesian, and performance input is constrained to these dimensions. The dimensionality reduction acts as funnel that does not provide access to the complete feature space of the corpus and may forsake the richness of performance input.

* The research leading to these results has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (Grant agreement No. 789825)

Fig. 1. A performer, wearing a sensor armband and engaging in a sound tracing study.

Our work builds upon a previous sound tracing study, where participants designed gestures to articulate time-varying sound using simple granular synthesis (Fig. 1). While sound tracing usually studies evoked gestural response [3], we extended traditional sound tracing to enable gesture-sound reproduction, and trained machine learning models to enable exploratory gestural performance by articulating expressive variations on the original sound. Participants were interested in exploring new timbres, but were limited by the provided corpus. The regression algorithm allowed them to scrub to different granular parameters in the stimulus sound, but not to a broader heterogeneous corpus. The study pointed out the potential for a more robust synthesis outlet for the gesture input regression model. Could we provide a corpus with sounds not in the original sound tracing stimulus? Could a regression model be harnessed to carry out feature mapping from the input domain (gesture) to the output domain (sound)?

We propose a system for exploring and performing with a multidimensional audio space using multimodal gesture sensing as the input. The input takes features extracted from electromyographic (EMG) and inertial sensors, and uses machine learning through regression modelling to create a contiguous gesture and motion space. EMG sensors on the forearm have demonstrated potential for expressive, multidimensional musical control, capturing small voltage variations associated with motions of the hand and fingers. The output target is generated via CBCS[11, 13], a technique which creates longer sounds by combining shorter sounds, called "units." A corpus of sounds is segmented into units which are catalogued by auditory features. Units can be recalled by query with a vector of those features. Our system allows musicians to quickly create an association between points and trajectories in a gesture feature space and units in a timbral feature space. The spaces can be explored and augmented together, interactively.

The paper is structured as follows. We first review related work on concatenative synthesis in performance. We then describe the proposed system, its archi-

tecture and technical implementation. Section 3 presents sound design strategies that address questions of corporal density to enable expressive performance, and the user workflow to associate gesture and CBCS sound via regression. In the discussion we provide a critical assessment of this approach and point out perspectives for future work before concluding.

2 Related Work

Aucouturier and Pachet [1] used concatenative sound synthesis to generate new musical pieces by recomposing segments of pre-existent pieces. They developed a constraint-satisfaction algorithm for controlling high-level properties like energy or continuity of the new track. They presented an example where a musician controls the system via MIDI, demonstrating an audio engine suitable for building real-time, interactive audio systems.

Stowell and Plumbley's [15] work focused on building associations between two differently distributed, unlabelled sets of timbre data. They succeeded in the implementation of a regression technique which learns the relations between a corpus of audio grains and input control data. In evaluating their system, they observed that such an approach provides a robust way of building trajectories between grains, and mapping these trajectories to input control parameters.

Schwarz et al. [14] used CataRT controlled through a 2D GUI in live performance taking five different approaches: (i) re-arranging the corpus in a different order than the original one, (ii) interaction with self-recorded live sound, (iii) composing by navigation of the corpus, (iv) cross-selection and interpolation between sound corpora, and (v) corpus-based orchestration by descriptor organisation. After performing in these five modes they concluded that CataRT empowers musicians to produce rich and complex sounds while maintaining precision in the gestural control of synthesis parameters. It presents itself as a blank canvas, without imposing upon the composer/performer any precise sonority.

In a later work [12], Schwarz et al. extended interaction modes and controllers (2D or 3D positional control, audio input) and concluded by stating the need for machine learning approaches in order to allow the user to explore a corpus by the use of XYZ-type input devices. They present gestural control of CataRT as an expressive and playful tool for improvised performance.

Savary et al. [9, 10] created *Dirty Tangible Interfaces*, a typology of user interfaces that favour the production of very rich and complex sounds using CataRT. Interfaces can be constantly evolved, irreversibly, by different performers at the same time. The interface is composed of a black box containing a camera and LED to illuminate a glass positioned above the camera, where users can position solid and liquid materials. Material topologies are detected by the camera, where a grey scale gradient is then converted into a depth map. This map is the projected onto a 3D reduction of the corpus space to trigger different grains.

Another example of gestural control of concatenative synthesis is the artistic project *Luna Park* by G. Beller [2]. He uses one accelerometer on top of each hand to estimate momentum variation, hit energy, and absolute position of the

hands. Two piezoelectric microphones responded to percussive patterns played in different zones of his body (one near the left hip and the other one near the right shoulder). Sensor data were then mapped to audio engine parameters to synthesise and interact with another performers recorded speech.

3 Methodology

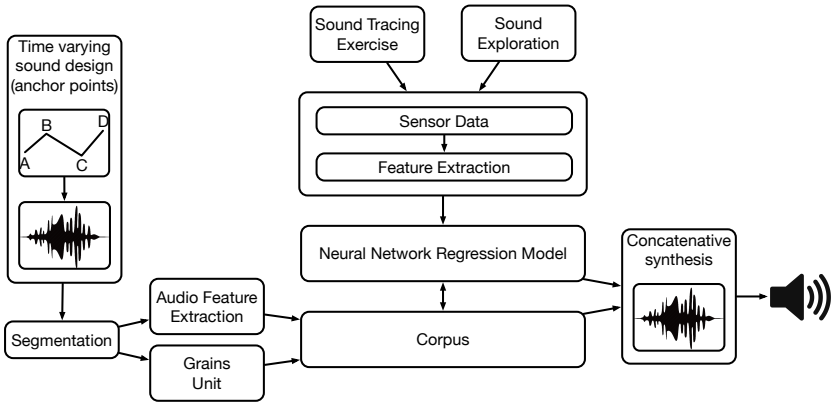


Fig. 2. System architecture diagram.

3.1 Implementation

We use a commercial sensor device[1], an armband worn on the forearm which packages eight electromyography (EMG) muscle sensors and an inertial measurement unit (IMU) for gross movement and orientation sensing, and transmits them over Bluetooth to a computer. We have also verified our approach with other biosensor packages, such as Plux's BITalino[2].

The software system is implemented in Cycling '74's Max[3]. We use the `myo`[4] object to capture raw EMG output of the sensors along with orientation quaternions from the on-board IMU to generate a multimodal feature vector representing the orientation, motion, and muscular state of the performer's forearm. Quaternions (x, y, z, and w: calculated by the device from accelerometer, gyroscope, and magnetometer data) give orientation, and we take the first-order difference between the current quaternion frame and the previous frame (x_d, y_d,

[1] `https://developerblog.myo.com/`
[2] `https://bitalino.com/`
[3] `https://cycling74.com/`
[4] `https://github.com/JulesFrancoise/myo-for-max`

z_d, w_d) to represent the current motion of the forearm. This is an important feature because hand gestures can be performed ballistically or in a more static fashion, causing different patterns of muscular activation even though the results are visually similar.

Raw EMG signals are intrinsically noisy, and we do not include them in our feature vector. Instead, we use a Bayesian[8] filter to probabilistically predict the amplitude envelope for each electrode in the armband. The sum of all eight amplitude envelopes is also included in the input feature vector, along with a new feature we have developed called "vector sum." Vector sum (Fig. 3) is a representation of the fact that the forearm muscles are situated around the arm in such a way that they can oppose or reinforce the action of other muscles. To calculate the vector sum, we model each electrode as representing a vector pointing away from the centre of a circle, evenly spaced every 45 degrees. The direction for each electrode vector does not change and the magnitude is proportional to the amplitude calculated by the Bayesian filter. The eight vectors are summed, and the resulting vector is related to the overall direction of force represented by all of the electrodes. When compared to the sum of all electrodes, the vector sum can distinguish gestures where muscles are opposing one another isometrically. This is an important feature, since joint movement might be minimal in such gestures but the subjective perception of effort is quite high. The vector sum is reported as a pair of Cartesian coordinates, which are better suited to regression than polar coordinates because they do not wrap around at zero degrees. See table 3.1 for a lists of the full gestural and timbral feature vectors. Where relevant, we took the average (μ) and standard deviation (σ) of each timbral feature over the whole audio unit.

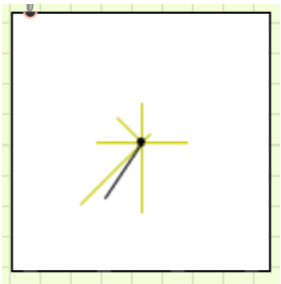


Fig. 3. An example vector sum, in black, drawn with the individual EMG vectors in yellow.

We implement machine learning in Max using an external object called `rapidmax`[7]. This object implements basic machine learning algorithms, such as multilayer perceptrons, k-nearest neighbour, and dynamic time warping, to allow Max users to quickly employ machine learning for regression or classification tasks. It is a Max wrapper around RapidLib [18], a C++ and JavaScript

library for creative, interactive machine learning applications in the style of Wekinator[4]. Specifically, we use a multilayer perceptron (MLP) neural network with one hidden layer to create models that perform regression based on user-provided training examples. This particular implementation uses a linear activation function on the output layer, allowing for model outputs that go beyond the numerical range of the provided examples, creating a larger and potentially more interesting generative space for aesthetic exploration. Training examples are created by associating inputs—gesture feature vectors—with outputs: vectors of timbral features. Performers can record example interactions, associating positions and gestures with sounds, to build an exploratory and performative gesture-timbre space.

Table 1. Input and output feature vectors for regression models

Input features(gesture)	Output features (timbre)
x	Duration
y	Frequency μ
z	Frequency σ
w	Energy μ
x_d	Energy σ
y_d	Periodicity μ
z_d	Periodicity σ
w_d	AC1 μ
EMG_1	AC1 σ
EMG_2	Loudness μ
EMG_3	Loudness σ
EMG_4	Centroid μ
EMG_5	Centroid σ
EMG_6	Spread μ
EMG_7	Spread σ
EMG_8	Skewness μ
EMG_{sum}	Skewness σ
$vectorSum_x$	Kurtosis μ
$vectorSum_y$	Kurtosis σ

The audio engine is implemented using MuBu[5] CataRT Max objects. We use the `mubu.process` object for segmentation and auditory feature analysis, `mubu.knn` for retrieval of the closest matching unit to a given set of auditory features, and the `mubu.concat` object for synthesising the unit once recalled in our workflow (see Section 3.3).

When a sound file is imported into MuBu, it is automatically segmented into units, either of a fixed length or determined by an onset detection algorithm (Fig. 4). A vector of auditory features (enumerated in table 3.1) is derived for each unit. These vectors of auditory features are associated with sensor feature

[5] `https://forumnet.ircam.fr/product/mubu-en/`

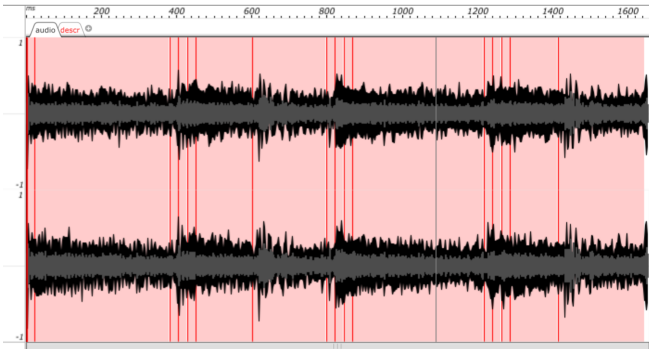


Fig. 4. A sound file imported into a MuBu buffer, using the onseg algorithm. Unit boundaries are shown as vertical, red lines. These lines would be equally spaced in chop mode. This display can be opened from the main gui window.

vectors to train a neural network, and roughly represent a high-dimensional timbral similarity space.

During playback, the amplitude and panning of the output is controlled by the "Amplitude Panner" (Fig. 5, upper right panel). The EMG sensors are divided into two groups and their amplitude envelopes are summed. The sum of each group is used to control the overall amplitude of the audio output in the left and right channels, respectively. When there is no muscular activation, both channels have near zero gain, giving the performer a natural method to make the instrument silent when they are not putting any energy into the system.

3.2 Sound & Gesture Design

In our previous study [16], we proposed four different approaches to designing gesture-timbre interaction based on a sound tracing exercise. In this system we revisit two of those approaches using our concatenative audio engine. Sound tracing is an exercise where a sound is given as a stimulus to study evoked gestural response [3]. Sound tracing has been used as a starting point for techniques of "mapping-by-demonstration" [5].

For that exercise, we used a general purpose software synthesizer, SCP by Manuel Poletti, controlled a breakpoint envelope-based playback system. We chose to design sounds that transition between four fixed anchor points with fixed synthesis parameters, primarily using SCP's granular synthesis engine. Envelopes interpolate between these fixed points. The temporal evolution of sound is captured as different states in the breakpoint editor whose envelopes run during playback. Any of the parameters can be assigned to breakpoint envelopes to be controlled during playback.

Users were asked to design a gesture that matched a pre-designed sound, and to train a regression model by associating data from that gesture with

the parameters of the sound. This created an exploratory space for performing variations on the source sound.

Our current work extends that activity with the use of CBCS. We go beyond the regression-based control of parametric synthesis from the previous study to create a mapping from gesture features to timbral features. This enables the user to perform the sound's corpus in real time, using variations on the original sound tracing gesture to articulate new sounds.

With this technique, we encounter potential problems of sparsity of the corpus feature space. There is no guarantee that there will be a unit that is closely related to the timbral features generated by the neural network in response to a given set of target gesture features. In order to address this, we added a step in our sound design method to fill the corpus with sound related to the original sound stimulus, but that had a wider range of timbral features. To do so, we went back to the original SCP sound authoring patch and replaced the source sample with a series of other sound samples. We then played the synthesis envelopes to generate time-varying sounds that followed the pitch/amplitude/parametric contour of the original stimulus. These timbral variants were recorded as separate audio files, imported into CataRT, and analysed. In this way, the corpus was enriched in a way that was directly related to the sound design of the original stimulus but had a greater diversity of timbral features, creating potential for more expressive variation in performance.

3.3 Workflow

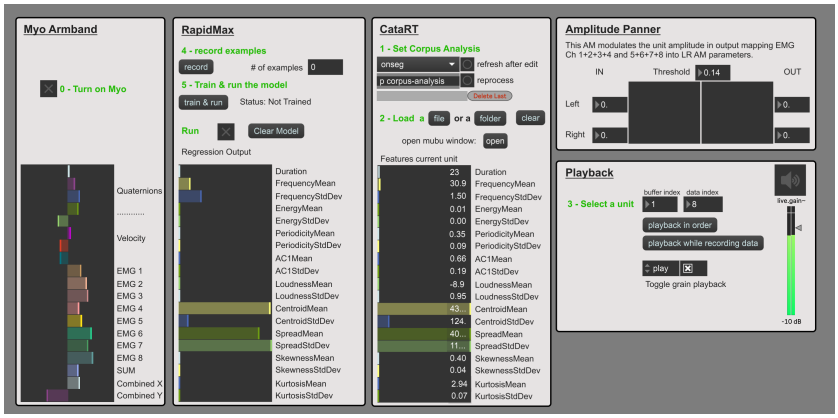


Fig. 5. Graphical-User Interface.

Fig. 5 shows the controls to interact with our system. At the beginning of a session, a performer activates the sensor armband using the toggle in the first column. The feature vector derived from the EMG/IMU sensors is displayed in

the same panel, allowing the performer to verify that the sensors are working as expected. They next select, in the CataRT column, the type of segmentation and analysis that will be performed on sound files imported into a corpus, choosing between onset-based segmentation and "chop," which divides the sound into equal-sized units. Parameters for these are available in a subpanel. The performer then imports individual sound files or folders of sounds into the corpus. Importing is cumulative; the whole corpus can be cleared with the clear button. It is also possible to re-analyse the current corpus using new parameters.

Once the sensors are activated and a corpus has been imported, segmented, and analysed, the gesture-timbre mapping process can begin. Performers may listen to individual units by selecting the buffer index (which file the unit is from) and data index (which unit in a file). The analysed timbral features associated with the selected unit are displayed by the multislider in the same column. In order to associate gestural features with the selected timbral features, they press the *record* button in the RapidMax column, which automatically captures 500ms of sensor data associated with the selected unit data.

Another way of interacting with audio units is to play an entire file from the first to the last unit. This enables the sound tracing workflow. Performers can listen to the whole sound and design the appropriate gesture to accompany that sound. Once that gesture has been designed, they can click on the *playback while recording data* button and then perform their gesture synchronously with sound playback. As each unit passes in order, the associated timbral data will be recorded in conjunction with the gestural data at that moment. This corresponds to the "whole gesture regression" mode from our previous work.

Once a set of potentially interesting example data has been recorded, users train the neural network by clicking the *train & run* button. When training has finished, the system enters run mode. At this point, incoming gestural data is sent to the trained regression model. This model outputs a vector of target timbral features that is sent to a k-nearest neighbours algorithm implemented in MuBu. That model outputs the unit buffer and data indices that most closely match the targeted timbral features. `mubu.concat` receives the indices and plays the requested unit. In this way, we have coupled the MLP model of gestural and timbral features to a k-NN search of timbrally defined units the corpus. This process allows musicians to explore the gesture-timbre space, and perform with it in real time.

4 Discussion

In lieu of a formal evaluation, in this section we reflect on the strengths and weaknesses of the system with a critical assessment of its affordances based on testing by the authors.

Concatenative synthesis is described in terms of a target that one tries to synthesise by navigating a corpus. In an audio-audio mosaicing task, the "target" is an example sound that one is trying to resynthesize with the corpus. In cases using interfaces for live controllers [9, 12], the mapping between gesture

and sound has, until now, taken place in a reduced dimension space. Two or three features are selected as pertinent and projected onto a graphical Cartesian representation. Our system does not require dimensionality reduction, and the number of input dimensions does not need to match the number output dimensions. This creates the disadvantage, however, of not being able to visualise the feature space. Schwarz in [12] finds seeing a reduced projection of the feature space convenient, but prefers to perform without it.

Schwarz describes exploratory performance as a DMI application of CBCS that distinguishes it from the more deterministic applications of speech synthesis or audio mosaicing [12]. He provides the example of improvised music where the performer uses an input device to explore a corpus, sometimes one that is being filled during the performance by live sampling another instrumentalist. This creates an element of surprise for the performer. Here we sought to create a system that enables timbral exploration, but that would be reproducible, and useful in compositional contexts where both sound and associated gesture can be designed *a priori*. In using our system, we were able to perform the sound tracing gesture to reconstruct the original sound. This shows that the generation of time-varying sound sources from our parametric synthesis programme were faithfully reproduced by CataRT in this playing mode.

Difficulties arose when we created gesture variations where the regression model started to "look for" units in the feature space that simply were not there, raising the problem of corpus sparseness. The sound design strategy to generate variants effectively filled the feature space. It was important that the feature space was filled not just with any sound, but with sound relevant to the original for which the gesture had been authored. By generating variants using different sound samples but that followed the same broad sonic morphology, we populate the corpus with units that were musically connected to the original but that were timbrally (and in terms of features) distinct. This creates a kind of hybrid between a homogeneous and heterogeneous corpus. It is heterogeneous in the diversity of sound at the performer's disposal, but remains musically coherent and homogeneous with the original sound/gesture design. This allowed expressive variation on a composed sound tracing gesture.

In order to support expressive performance we need to create gesture-timbre spaces that maximise sonic diversity. When a performer navigates through gesture space, the outcome is more rich and expressive if a diverse range of individual units are activated. The nuances of gesture become sonically meaningful if that gestural trajectory has a fine-grained sonic result. This is not always the result of the proposed workflow, especially in the case where the user chooses individual units and associates them with specific gestural input. It is, again, a problem of sparseness; the distribution of units in a high-dimensional space is not usually even. In a typical corpus, there will be areas with large clusters of units and other units that are relatively isolated. When musicians choose individual units to use for gesture mapping, there is a tendency to choose the units that have the most character. These units are often outliers. In a good outcome, outlier units represent the edges of the timbral space of the corpus. In this case, a regression

between units on different edges of the space will activate a wide range of intermediate units. However, it is also possible that a gestural path between two interesting units does not pass near any other units in the corpus. In that case, the resulting space performs more like a classifier, allowing the performer to play one unit or another without any transitional material between them. It would be helpful to give users an idea about where the potentially interesting parts of a corpus lie. It might also be possible to automatically present performers with units that represent extreme points of the timbral space, or areas where gesture mapping might yield interesting results.

Future work to address varying sparseness and density of the corpus feature space maybe be in dynamic focus on areas more likely to have sound. Schwarz in [12] uses Delaunay triangulation to evenly redistribute the three dimensional projection of the corpus in his tablet based performance interface. This operation would be more difficult in a higher dimensional space. One recent development in the CataRT community has been the exploration of using self-organising maps to create a more even distribution of features in the data space [6].

Another potentially interesting way to use multidimensional gesture-timbre mapping is to generate feature mapping using one corpus of sounds, and then either augment that corpus or change to an entirely different corpus – moving the timbral trajectory of a gesture space into a new set of sounds. This can be fruitful when units in the new corpus intersect with the existing gesture space, but it is difficult to give users an idea about whether or not that will be the case. One idea we are exploring is "transposing" a trajectory in timbral feature space so that it intersects with the highest number of units in a new corpus. This could be accomplished using machine learning techniques, such as dynamic time warping, to calculate the "cost" of different ways to match a specific trajectory to a given set of units, and find the optimal transposition. We know from Wessel's seminal work on timbre spaces [17] that transposition in a low dimensional timbre space is perceptually relevant. Automatically generating these transpositions, or suggesting multiple possible transpositions, has the potential to generate novel musical phrases that are perceptually connected to the original training inputs.

5 Conclusions

We have presented a system that combines regression-based machine learning with corpus-based concatenative synthesis. We extend a previous study where a sound tracing workflow was used to design gestures to articulate time varying sounds. Gesture input from EMG and IMU sensors generated multidimensional targets and are associated with specific points in a high-dimensional timbral feature space in order to train a neural network. Using this workflow, we were able to reproduce original sound tracings. By populating the corpus with related, but timbrally diverse grains, we increased the corporal density to enable expressive variation on the original gesture. This workflow demonstrates the use of real-time, interactive machine learning with CataRT and creates a multidimensional feature mapping linking gesture to sound synthesis.

References

1. J.-J. Aucouturier and F. Pachet, *Jamming with Plunderphonics: Interactive concatenative synthesis of music*, Journal of New Music Research, 35 (2006), pp. 35–50.
2. G. Beller, *Gestural control of real time speech synthesis in lunapark*, in Proceedings of Sound Music Computing Conference, SMC, Padova, Italy, 2011.
3. B. Caramiaux, F. Bevilacqua, and N. Schnell, *Towards a gesture-sound cross-modal analysis*, in Gesture in Embodied Communication and Human-Computer Interaction, Berlin, Heidelberg, 2010, pp. 158–170.
4. R. Fiebrink and P. R. Cook, *The wekinator: a system for real-time, interactive machine learning in music*, in Proceedings of The International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, 2010.
5. J. Françoise, *Motion-sound mapping by demonstration*, PhD thesis, UPMC, 2015.
6. J. Margraf, *Masters thesis*, TU Berlin, (2019).
7. S. T. Parke-Wolfe, H. Scurto, and R. Fiebrink, *Sound control: Supporting custom musical interface design for children with disabilities*, in Proceedings of the International Conference on New Interfaces for Musical Expression, NIME'19, Porto Alegre, Brazil, 2019.
8. T. D. Sanger, *Bayesian filtering of myoelectric signals*, Journal of neurophysiology, 97 (2007), pp. 1839–1845.
9. M. Savary, D. Schwarz, and D. Pellerin, *Dirti —dirty tangible interfaces*, in Proceedings of the International Conference on New Interfaces for Musical Expression, NIME'12, Ann Arbor, Michigan, 2012.
10. M. Savary, D. Schwarz, D. Pellerin, F. Massin, C. Jacquemin, and R. Cahen, *Dirty tangible interfaces: Expressive control of computers with true grit*, in CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, Paris, France, 2013, ACM, pp. 2991–2994.
11. D. Schwarz, *Concatenative sound synthesis: The early years*, Journal of New Music Research, 35 (2006), pp. 3–22.
12. D. Schwarz, *The sound space as musical instrument: Playing corpus-based concatenative synthesis*, in Proceedings of the International Conference on New Interfaces for Musical Expression, NIME'12, Ann Arbor, Michigan, 2012.
13. D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton, *Real-Time Corpus-Based Concatenative Synthesis with CataRT*, in 9th International Conference on Digital Audio Effects, DAFx 19, Montreal, Canada, 2006, pp. 279–282.
14. D. Schwarz, R. Cahen, and S. Britton, *Principles and applications of interactive corpus-based concatenative synthesis*, in Journes d'Informatique Musicale, JIM, Albi, France, 2008.
15. D. Stowell and M. D. Pumbley, *Timbre remapping through a regression-tree trchnique*, in Proceedings of the Sound Music Computing Confetence, SMC, 2010.
16. A. Tanaka, B. Di Donato, and M. Zbyszyński, *Designing gestures for continuous sonic interaction*, in Proceedings of the International Conference on New Interfaces for Musical Expression, NIME'19, Porto Alegre, Brazil, 2019.
17. D. L. Wessel, *Timbre space as a musical control structure*, Computer music journal, (1979), pp. 45–52.
18. M. Zbyszyński, M. Grierson, and M. Yee-King, *Rapid prototyping of new instruments with codecircle*, in Proceedings of the International Conference on New Interfaces for Musical Expression, Copenhagen, Denmark, 2017, pp. 227–230.

Beyond the Semantic Differential: Timbre Semantics as Crossmodal Correspondences

Charalampos Saitis

Centre for Digital Music, Queen Mary University of London, London, UK
c.saitis@qmul.ac.uk

Abstract. This position paper argues that a systematic study of crossmodal correspondences between timbral dimensions of sound and perceptual dimensions of other sensory modalities (e.g., brightness, fullness, roughness, sweetness) can offer a new way of addressing old questions about the perceptual and cognitive mechanisms of timbre semantics, while the latter can provide a test case for better understanding crossmodal correspondences and human semantic processing in general. Furthermore, a systematic investigation of auditory-nonauditory crossmodal correspondences necessitates auditory stimuli that can be intuitively controlled along intrinsic continuous dimensions of timbre, and the collection of behavioural data from appropriate tasks that extend beyond the semantic differential paradigm.

Keywords: Timbre semantics · Crossmodal correspondences · Sound morphing · Conceptual metaphor · Acousmatic sound · Sound quality

1 Introduction

Timbre is a perceptual property of auditory objects, encompassing a complex set of attributes collectively contributing to the inference of a sound's source but also acting as qualia [26]. Bright, rough, hollow, nasal, or velvety are just a few examples of the linguistic descriptions engaged by music professionals and other expert and naïve listeners when they communicate timbral qualities of sounds. These metaphorical descriptions are central to conceptualizing timbre by allowing to make sense of its perceptual representation through indexing other, more commonly shared embodied semantic representations [31, 18, 21]. As such, a relation must exist between the acoustical properties of a sound that give rise to a timbral nuance and its semantic description.

Nonauditory sensory attributes of timbre (e.g., bright, rough, hollow) particularly exemplify a more ubiquitous aspect of human cognition known as *crossmodal correspondences*: people systematically associate between sensory experiences presented in different modalities (e.g., between timbre and touch [19]) or within the same modality (e.g., between pitch, timbre, and loudness [10]). Studying crossmodal correspondences between timbral dimensions of sound and perceptual dimensions of other sensory modalities can therefore offer a new way of addressing old questions about the perceptual and cognitive processes underlying the metaphorical ways we talk about sound.

At the same time, timbre and the crossmodal metaphors that dominate its conceptualization can provide a test case for better understanding crossmodal correspondences. While there is ample evidence of consistently regular mappings between modalities, the knowledge of both the psychophysics and higher cognitive processes that govern those mappings is still rather narrow. In the case of sound, there is a growing body of studies documenting the behavior of associations between pitch and other modalities (see [29] for a review) but similar research on timbre, including vowels, is still very limited [9, 32, 4, 27, 1, 16, 5].

Accordingly, this contribution is both a position paper and a work in progress. Following a short overview of research in timbre semantics in Sect. 2, Sect. 3 discusses important aspects that should be considered when studying auditory-nonauditory crossmodal correspondences, with a special emphasis on the design and implementation of appropriate auditory stimuli. To this end, an acousmatic and sound morphing approach to timbre analysis-synthesis is outlined, based on well-established methods from psychoacoustics and sound synthesis.

2 Crossmodal Metaphors in Timbre Cognition

In his empirical ethnographic research on the management of discourse about sound between music professionals in the United States, Porcello [15] identified five strategies that are common among producers and engineers in how they talk about timbre: *spoken/sung vocal imitations* of timbral characteristics; *lexical onomatopoeic metaphors*; *pure metaphor* (i.e., non-onomatopoeic, generally referencing other sensory modalities or abstract concepts); *association* (citing styles of music, musicians, producers, etc.); and *evaluation* (judgments of aesthetic and emotional value). Wallmark [30] similarly categorized verbal descriptions of instrumental timbre collected across 11 orchestration treatises according to: *affect* (emotion and aesthetics); *matter* (features of physical matter); *crossmodal correspondence* (CMC; borrowed from other senses); *mimesis* (sonic resemblance); *action* (physical actions and qualities of movement); *acoustics* (auditory and spatial sound terminology); and *onomatopoeia* (phonetic resemblance).

Porcello [15] especially advances a distinction between vocal imitations and onomatopoeias on the one hand, which he denotes as "sonic iconicity", and the "pure iconicity" of metaphors originating in nonauditory sensory experiences or abstract concepts on the other. Wallmark [31] argues that what Porcello singles out as *pure metaphor*, which could be seen as encompassing *matter*, *crossmodal correspondence*, *mimesis*, *action*, and *acoustics* in Wallmark's categorization, is central to the process of conceptualizing timbre and not simply a matter of linguistic convention. Timbre, Wallmark proposes, appears to be experienced through three conceptual metaphors grounded in corporeal experiences: INSTRUMENTS ARE VOICES, SOUND IS MATERIAL, and NOISE IS FRICTION. In other words, timbre can be meaningfully experienced with reference to articulatory motor representations of human and non-human voice production (e.g., nasal, hissing, raspy) and as a material object that can be seen, touched, and even tasted (e.g., dark, rough, sweet).

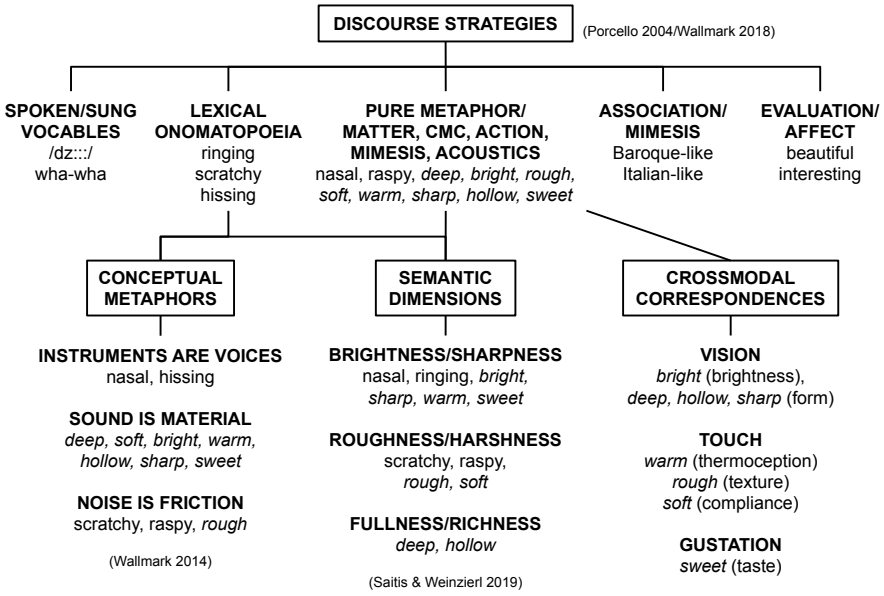


Fig. 1. The discourse and semantics of timbre

The most popular approach to the study of timbre attempts to quantify the ways in which sounds are perceived to differ through multidimensional scaling of pairwise dissimilarity ratings [26]. While this method has been effective in exploring the perceptual representation of timbre, it is not designed to investigate its semantic dimensions. These are more commonly studied by means of factor analysis of ratings along *semantic differentials* [14]. The latter are typically constructed either by two opposing descriptive adjectives such as "bright–dull" or by an adjective and its negation as in "bright–not bright". Semantic differential studies of timbre aim to identify the few salient semantic substrates of its diverse discourse that can yield consistent and differentiating responses to different sounds along with their acoustic correlates.

Von Bismark [2] used synthetic spectra that mimicked vowels and instruments and empirically derived verbal scales (in German) suitable for describing such timbres (as opposed to a priori selection by the experimenter) and settled for a four-dimensional semantic space for timbre. The first dimension was defined by the differential scale *dull–sharp*, explained almost half of the total variance in the data, and correlated well with the spectral centroid. In an English experiment taking up some of Bismarck's verbal scales but using dyads played from different wind instruments [7], it was found that dull–sharp ratings were less stable, likely because sharp in English refers more often to pitch than to timbre. Evidence from all subsequent studies in English (and in several other languages) corroborate that the salient dimension of timbre related to spectral energy distribution and

concentration of energy in higher frequency bands is captured by the pair of polar adjectives *dull–bright*.

The other dimensions found by von Bismarck were *compact–scattered*, *full–empty*, and *colorful–colorless*. Today most semantic differential designs will yield a single dimension of *fullness* that encompasses all such timbral impressions as well as a third common dimension of *roughness* or *harshness* related to narrow harmonic intervals, rapid amplitude fluctuations, but also to excessive, unpleasant high-frequency energy [33]. These results are generally also corroborated by investigations based on multidimensional scaling of adjective dissimilarities [12] and psycholinguistic analyses of verbalization tasks [18, 20, 17]. The boundaries between these dimensions are sometimes blurred, while different types of timbres or scenarios of timbre perception evoke semantic dimensions that are specific to each case (see [21] for a comprehensive review).

The semantic differential method has been instrumental in advancing our understanding of timbre semantics [2, 7, 33]. However, the underlying cognitive mechanisms remain largely unknown and underexplored, especially when one looks beyond the case of acoustic musical instruments. In viewing timbre semantics through the lens of crossmodal correspondences, questions about the perceptual and cognitive mechanisms underlying both phenomena can thus be reconsidered: What timbral properties of sound evoke the analogous impression as viewing a rounded form, touching a smooth surface, or tasting a sweet edible? How are auditory-nonauditory crossmodal mappings structured across modalities? Are attributes of different sensory experiences (e.g., a rounded form, a smooth surface, and a sweet edible) mapped to similar or distinct timbres?

Addressing these questions requires a systematic exploration of crossmodal correspondences between timbral dimensions of sound and perceptual dimensions of other modalities, including the collection of behavioural data from appropriate tasks that extend beyond the semantic differential paradigm.

3 An Acousmatic and Sound Morphing Framework for Timbre Analysis-Synthesis

Previous work on timbre-nonauditory crossmodal correspondences has relied on recorded notes from acoustic musical instruments, which may implicate source-cause categories rather than continuous sensory dimensions [25]. In their study of crossmodal associations between sounds and tastes, Simner and colleagues [27] used formant synthesis to create four vowel quality continua (first formant, second formant, voice discontinuity, spectral balance). Listeners selected preferred positions along the continua to accompany each of the basic tastes of sweet, sour, bitter, and salty, each received at two different concentrations. It was found that mappings of sour, bitter, and salty generally patterned with each other, while sweet patterned away from all three other tastes. Indeed, in taste perception humans often confuse sour with salty and/or bitter, and occasionally salty with bitter, but they always discriminate between sweet and all other tastes [13]. It was further observed that, when participants matched sounds of low to mid to

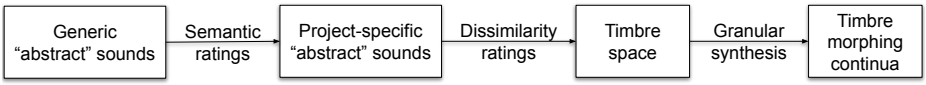


Fig. 2. Proposed analysis-synthesis approach to develop timbre morphing continua based on dissimilarity ratings of semantically relevant acousmatic sounds

high frequency to different tastes, a corresponding hierarchy of sweet to bitter to sour emerged.

Expanding on these ideas, a timbre analysis-synthesis framework is proposed to design “abstract” sounds that lack readily available source/cause associations such as those found in instrumental and vocal timbre classes, and which can be *morphed* along perceptual dimensions of timbre (Fig. 2). Considering two sounds placed at either end of one timbral dimension, sound morphing techniques will allow to create new sounds whose timbres lie perceptually between the timbres of the two original sounds. *Timbre morphing continua* can thus give listeners a simple and intuitive way to navigate among available timbres, while using “abstract” sound morphologies as the source material will help ensure control of *intrinsic* dimensions of timbre.

A useful way of collecting such sounds is through electroacoustic music, where sound identities appear intentionally obscured or unconnected to their source, and which have been shown to activate crossmodal associations of the type proposed to investigate here [23, 6, 8]. Researchers at the PRISM laboratory in Marseille have previously compiled a collection of 200 electroacoustic sounds representative of the nine “balanced sounds” classes of Schaeffer’s typology of *acousmatic* sounds—sounds experienced by attending to their intrinsic morphology and not to their physical cause [11]. These are based on three profiles of temporal energy envelope (continuous, impulse, iterative) and three profiles of spectral content (tonal, complex pitch, varying pitch) [22]. This classification system offers an objective tool to obtain a sound corpus representative of most sound morphologies.

However, sounds with iterative envelopes and/or varying pitch may not be suitable to study intrinsic qualia aspects of timbre. For example, the second type tends to evoke motion [11]. Accordingly, starting from these 200 samples (made available courtesy of PRISM), only those sounds that are representative of the four {continuous, impulse}×{tonal, complex pitch} classes of Schaeffer’s typology will be selected through an informal listening protocol as the generic sound material for the proposed timbre analysis and synthesis. Different durations and pitches within the generic sound corpus will be retained initially.

The most commonly used crossmodal sensory attributes of timbre typically relate to the modalities of vision (*bright, dark, dull, deep, thick, thin, hollow, full, round, rounded, sharp*), touch (*soft, hard, smooth, rough, warm*), and gustation (*sweet*). Each sound will be rated along corresponding verbal scales to quantify the extent to which it evokes each of these attributes. Listeners will be advised to use a scale only when they feel that the corresponding attribute is evoked by the given sound. If no scales have been used in a trial, raters will see a warning

message before proceeding to the next sound, asking them to confirm their no-scale choice by means of a check box. By looking at scale-based interindividual consistency and stimulus-based mean attribute magnitudes, only those sounds that have been judged as evocative of *all* attributes will be considered for further analysis. This will help obtain a reduced sound set that is optimum in view of the hypothesis (i.e., in view of the investigated crossmodal associations).

The reduced sounds will be subjected to dissimilarity ratings by a different group of listeners. While sounds of the generic corpus will be allowed to vary in both duration and pitch for the semantic ratings, these two variables, which covary with timbre [26], will need to be equalized as much as possible in the reduced sound set for the dissimilarity ratings. Through multidimensional scaling (MDS) of dissimilarities and audio content analysis of the acoustic signals [3], a timbre space will be established. Because the tested sounds will be selected to fit morphological criteria, and based on previous work in electroacoustic sounds [6, 8], the timbre space is expected to have between three and five dimensions related to spectral energy distribution (*location* of energy concentrations in the frequency axis), spectral bandwidth (*extensity* in the frequency axis), and spectral *density* (energy distribution in relation to spectral bandwidth) [22, 28].

Available sound morphing techniques and software mostly allow to create generic hybrids of the two end sounds without attending to specific timbral dimensions. Here it is proposed to use granular synthesis, where desired sounds are created by reorganizing tiny snippets of other sounds called grains, to create continua between sounds placed at the corner marks of each dimension of the obtained timbre space: at each end of a continuum there will be only grains from the corresponding sounds; along the continuum, intermediate timbres will result from continuously morphing between pairs of similar grains and/or the global spectral envelopes; the nature of morphing will be determined according to the modelled timbral dimension. For instance, redistributing the energy in spectrum A so that it resembles that of spectrum B can be thought of as shifting the formant locations of A to those of B [24]. A granular representation appears better suited for the kind of acousmatic sounds proposed here as the source material [8]. Morphing based on sinusoidal modelling would be more appropriate for sounds with prominent oscillatory modes, such as musical instruments [3].

4 Conclusion and Work in Progress

It is argued here that a systematic investigation of crossmodal correspondences between between timbral and nonauditory perceptual dimensions (e.g., brightness, roughness, sweetness) can offer a new perspective into the study of sensory semantics, which are central in human behaviour as they include important aspects of the conceptual knowledge acquired about the world. Forming the first step towards such an investigation, an acousmatic and sound morphing framework for timbre analysis-synthesis is currently under development to obtain morphing continua for intuitive control of intrinsic dimensions of timbre.

Subsequently, in crossmodal matching experiments also currently in the pipeline, auditory-nonauditory crossmodal correspondences will be quantified in terms of shared mappings between physical or virtual visual/tactile/gustatory stimuli varying along their modality-specific dimensions and synthesized sounds varying along the timbre morphing continua. The latter will remain the same across the different crossmodal matching designs in order to always investigate the same timbral dimensions.

Acknowledgments. The author wishes to thank Sølvi Ystad, Christine Cuskley, and Christoph Reuter for their valuable and constructive suggestions during the planning and development of this research work. I am particularly grateful to Sølvi for sharing acousmatic sound materials from previous work at PRISM.

References

1. Adeli, M., Rouat, J., Molotchnikoff, S.: Audiovisual correspondence between musical timbre and visual shapes. Front. Hum. Neurosci. **8**, 352 (2014)
2. von Bismarck, G.: Timbre of steady tones: A factorial investigation of its verbal attributes. Acustica **30**, 146–159 (1974)
3. Caetano, M., Saitis, C., Siedenburg, K.: Audio content descriptors of timbre. In: K. Siedenburg, C. Saitis, S. McAdams, A.N. Popper, R.R. Fay (eds.) Timbre: Acoustics, Perception, and Cognition, pp. 297–333. Springer, Cham (2019)
4. Crisinel, A.S., Spence, C.: As bitter as a trombone: Synesthetic correspondences in nonsynesthetes between tastes/flavors and musical notes. Atten. Percept. Psychophys. **72**(7), 1994–2002 (2010)
5. Cuskley, C., Dingemanse, M., Kirby, S., van Leeuwen, T.M.: Cross-modal associations and synaesthesia: Categorical perception and structure in vowel-colour mappings in a large online sample. Behav. Res. Methods (published online) (2019)
6. Grill, T.: Perceptually informed organization of textural sounds. Ph.D. thesis, University of Music and Performing Arts Graz, Graz, Austria (2012)
7. Kendall, R.A., Carterette, E.C.: Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. Music Percept. **10**, 445–468 (1993)
8. Lembke, S.A.: Hearing triangles: Perceptual clarity, opacity, and symmetry of spectrotemporal sound shapes. J. Acoust. Soc. Am. **144**, 608–619 (2018)
9. Marks, L.E.: On colored-hearing synesthesia: cross-modal translations of sensory dimensions. Psychol. Bull. **82**, 303 (1975)
10. Melara, R.D., Marks, L.E.: Interaction among auditory dimensions: Timbre, pitch, and loudness. Percept. Psychophys. **48**, 169–178 (1990)
11. Merer, A., Aramaki, M., Ystad, S., Kronland-Martinet, R.: Perceptual characterization of motion evoked by sounds for synthesis control purposes. ACM Trans. Appl. Percept. **10**, 1–24 (2013)
12. Moravec, O., Štěpánek, J.: Verbal description of musical sound timbre in czech language. In: Proc. 3rd Stockholm Music Acoust. Conf. (SMAC 03), vol. II, pp. 643–645. Stocholm, Sweden (2003)
13. Mueller, C.A., Pintscher, K., Renner, B.: Clinical test of gustatory function including umami taste. Ann. Otol. Rhinol. Laryngol. **120**, 358–362 (2011)
14. Osgood, C.E.: The nature and measurement of meaning. Psychol. Bull. **49**(3), 197–237 (1952)

15. Porcello, T.: Speaking of sound: language and the professionalization of sound-recording engineers. Soc. Studies Sci. **34**(5), 733–758 (2004)
16. Reuter, C., Jewanski, J., Saitis, C., Czedik-Eysenberg, I., Siddiq, S., Oehler, M.: Colors and timbres – consistent color-timbre mappings at non-synesthetic individuals. In: Proceedings of the 34. Jahrestagung der Deutschen Gesellschaft für Musikpsychologie: Musik im audiovisuellen Kontext. Gießen, Germany (2018)
17. Saitis, C., Fritz, C., Scavone, G.P.: Sounds like melted chocolate: how musicians conceptualize violin sound richness. In: Proceedings of the 2019 International Symposium on Musical Acoustics. Detmold, Germany (2019)
18. Saitis, C., Fritz, C., Scavone, G.P., Guastavino, C., Dubois, D.: Perceptual evaluation of violins: A psycholinguistic analysis of preference verbal descriptions by experienced musicians. J. Acoust. Soc. Am. **141**(4), 2746–2757 (2017)
19. Saitis, C., Järveläinen, H., Fritz, C.: The role of haptic cues in musical instrument quality perception. In: S. Papetti, C. Saitis (eds.) Musical Haptics, pp. 73–93. Springer, Cham (2018)
20. Saitis, C., Weinzierl, S.: Concepts of timbre emerging from musician linguistic expressions. J. Acoust. Soc. Am. **141**, 3799 (2017)
21. Saitis, C., Weinzierl, S.: The semantics of timbre. In: K. Siedenburg, C. Saitis, S. McAdams, A.N. Popper, R.R. Fay (eds.) Timbre: Acoustics, Perception, and Cognition, pp. 119–149. Springer, Cham (2019)
22. Schaeffer, P.: Traité des objets musicaux: essai interdisciplines. Editions du Seuil, Paris (1966), English edition: Schaeffer, P. (2017). *Treatise on musical objects: an essay across disciplines* (trans: North, C., Dack, J.; University of California Press, Oakland)
23. Schön, D., Ystad, S., Kronland-Martinet, R., Besson, M.: The evocative power of sounds: Conceptual priming between words and nonverbal sounds. J. Cogn. Neurosci. **22**, 1026–1035 (2009)
24. Siddiq, S.: Morphing of granular sounds. In: Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15), pp. 4–11 (2015)
25. Siedenburg, K., Jones-Mollerup, K., McAdams, S.: Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. Front. Psychol. **6**, 1977 (2016)
26. Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay, R.R. (eds.): Timbre: Acoustics, Perception, and Cognition. Springer, Cham (2019)
27. Simner, J., Cuskley, C., Kirby, S.: What sound does that taste? Cross-modal mappings across gustation and audition. Perception **39**(4), 553–569 (2010)
28. Smalley, D.: Spectromorphology: explaining sound-shapes. Organised Sound **2**, 107–126 (1997)
29. Walker, P.: Cross-sensory correspondences: A theoretical framework and their relevance to music. Psychomusicology **26**, 103–116 (2016)
30. Wallmark, Z.: A corpus analysis of timbre semantics in orchestration treatises. Psychol. Music (published online) (2018)
31. Wallmark, Z.T.: Appraising timbre: Embodiment and affect at the threshold of music and noise. Ph.D. thesis, University of California, Los Angeles (2014)
32. Ward, J., Huckstep, B., Tsakanikos, E.: Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all? Cortex **42**, 264–280 (2006)
33. Zacharakis, A., Pastiadis, K., Reiss, J.D.: An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. Music Percept. **31**, 339–358 (2014)

Generative Grammar Based on Arithmetic Operations for Realtime Composition

Guido Kramann[1]

Brandenburg University of Applied Sciences
kramann@th-brandenburg.de

Abstract. Mathematical sequences in $\mathbb{N}_0$ are regarded as time series. By repeatedly applying arithmetic operations to each of their elements, the sequences are metamorphised and finally transformed into sounds by an interpretation algorithm. The efficiency of this method as a composition method is demonstrated by explicit examples. In principle, this method also offers laypersons the possibility of composing. In this context it will be discussed how well and under what kind of conditions the compositional results can be predicted and thus can be deliberately planned by the user. On the way to assessing this, Edmund Husserl's concept of "fulfillment chains" provides a good starting point. Finally, the computer-based board game MODULO is presented. Based on the here introduced generative grammar, MODULO converts the respective game situation directly into sound events. In MODULO, the players behave consistent to the gaming-rules and do not care about the evolving musical structure. In this respect, MODULO represents an alternative draft to a reasonable and common use of the symbols of the grammar in which the user anticipates the musical result.

Keywords: algorithmic composition, phenomenology, arithmetic operations, realtime composition, live coding, Edmund Husserl, notation system

1 Introduction

This thesis deals with a generative process in the field of real-time composition, which is essentially based on the fact that different arithmetic operations are repeatedly applied to the elements of a mathematical sequence. In the following, this basic procedure shall be abbreviated as **AOG** (**A**rithmetic-**O**peration-**G**rammar).

"Every human is a composer" – with this casual modification of a saying by Joseph Beuys I would like to express that generative composition processes basically open up the possibility that even people with little knowledge of music theory can compose, since in the sense of Chomsky's division of generative grammars those of level 3 – the one presented here is one of this kind – help to produce exclusively meaningful/wellformed musical structures [2], [3].

Typically, generative methods of composition are judged from the point of view of what kind of structures they produce and if so what relation they have to music [17].

However, in the second part of this work, the actual process is discussed under another aspect, namely the extent to which the generative process chain can also be mapped in the mind of a person who produces it, especially with regard to its possible use as a composition aid for laypersons. To even consider taking such a direction is motivated by the fact that the overall procedure presented here works in such a way that the process of generating the composition from its symbolic representation is straightforward, without the need for automatic corrections or optimizations of the linear or harmonic structure. At least this ensures a relative transparency of the generating process.

But first the actual procedure is described in detail both theoretically and in examples and its special characteristics are analyzed:

2 A 3rd order generative grammar based on arithmetic operations applied to mathematical sequences

The overall shape of a sequence such as $a_{i+1} = a_i + 1$ (identity on natural numbers), or $a_{i+1} = a_i + a_{i-1}$ (fibonacci sequence) is to be changed by applying an arithmetic operation to each of its sequence members. This can be repeated on the resulting sequence with another operation, and so on. One gets a metamorphosis of sequences which have a close structural relation to each other.

For musical use, from now on all sequences are to be understood as time sequences which deliver their values in a fixed time interval ΔT within a real-time composition process.

Restrictively, initially only $id(\mathbb{N}_0) = \{0, 1, 2, 3, 4, ...\}$ is to be used as a source or time base, to which the arithmetic operations are subsequently applied.

$\mathbb{N}_0$ is also the permitted number range. So that this number range is never left, a filter is set after the execution of any operation, in which the decimal places are truncated and values smaller than zero are set to zero. In table 1 some operators are suggested to be used for this grammar. There the symbols used for the operations and their meaning are shown together with an example. In addition, it is shown here how the corresponding operator is represented in the game "MODULO" introduced at the end of this presentation.

The operators proposed here go a little beyond of what is common in arithmetic. In order to understand the table, the operators $\neq$, $==$, $\nmid$, $|$ should also be regarded as a type of filter that allows a number to pass when the condition meant is fulfilled.

symbol	symbol in MODULO	meaning	example
$+$	$+$	addition	$\{0,1,2,3,4\}+3=\{3,4,5,6,7\}$
$-$	$-$	subtraction	$\{0,1,2,3,4\}-3=\{0,0,0,0,1\}$
$\cdot$	$\cdot$	multiplication	$\{0,1,2,3,4\}\cdot 2=\{0,2,4,6,8\}$
$\neq$	$++$	not equal	$\{0,1,2,3,4\}\neq 3=\{0,1,2,0,4\}$
$==$	$--$	identity	$\{0,1,2,3,4\}==3=\{0,0,0,3,0\}$
$\div$	$\cdot\cdot$	division	$\{0,1,2,3,4\}\div 2=\{0,0,1,1,2\}$
$\nmid$	$+++$	does not divide	$\{0,1,2,3,4\}\nmid 2=\{0,1,0,3,0\}$
$\equiv$	$---$	modulo	$\{0,1,2,3,4\}\equiv 3=\{0,1,2,0,1\}$
$\mid$	$\cdots$	true divider	$\{0,1,2,3,4\}\mid 2=\{0,0,2,0,4\}$

Table 1: Used operators with examples.

2.1 Sound generation on the basis of a mathematical sequence

For sound generation, $id(\mathbb{N}_0)$ is now executed as a counting process with constant speed. The introduced grammar makes it easy to gradually increase the complexity of simple structures by adding an operation. Thus, adding a symbol on the symbol level typically results in an increase of complexity at the score level.

Each intermediate result of the successive operations is used in parallel for the sound generation. Thus, the members of each resulting sequence, including those resulting from the intermediate operations, are regarded as divisors of the base number b, with for example $b = 2520 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 5 \cdot 7$.

Whenever the divisibility is actually given and a number between for example 55 and 1760 (A1 and A6) comes out, this is understood as frequency, which is then mapped in the best possible way to the equally tempered scale, so that this tone can then be played in real time e.g. as a piano tone by a sequencer. This mechanism plays the role of a filter that suppresses pitches that have a too large harmonic difference to the overall structure.

2.2 "$\equiv 7 \cdot 5 \equiv 3 \cdot 5$" – a simple composition as an example

"$\equiv 7 \cdot 5 \equiv 3 \cdot 5$" is meant as a symbolic representation of a tiny composition (for sound and complete score see [7]). As it is a convention to apply all operators to $id(\mathbb{N}_0)$ first this information can be neglected in the symbolic representation. As an additional convention one operation is applied after the other with a time delay of twelve times ΔT which can be interpreted as two three-four time bars. Figure 1 shows how the unfolding of this composition could take place starting from the symbolic representation. Obviously, as we go through the successive stages of the unfolding process, there is a steady increase of information and complexity in the resulting structure.

2.3 Analysis of the Musical Structure

At first glance the resulting musical structure seems to be very similar to (repetitive) minimal music. This will be analysed in more detail here.

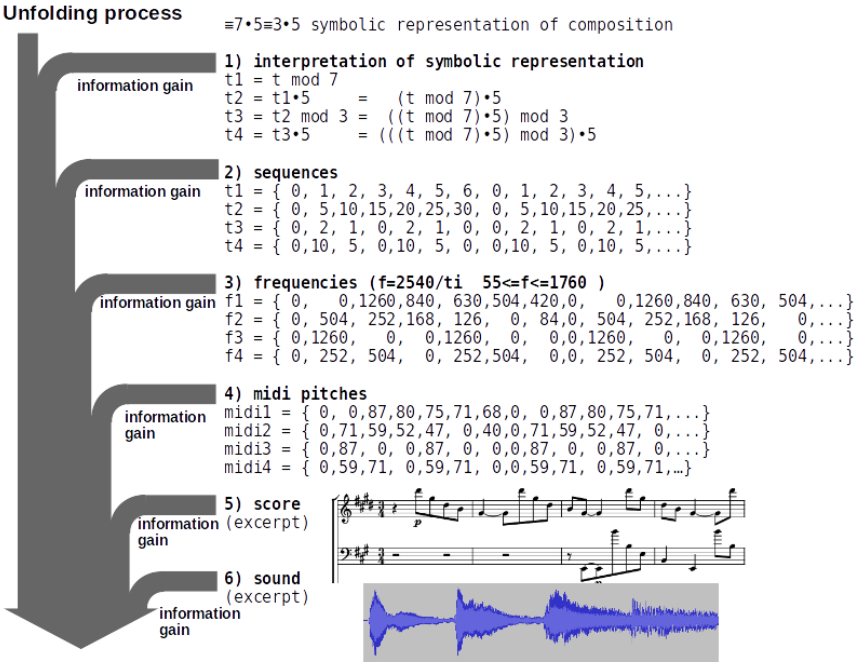


Fig. 1. Unfolding process from symbolic representation to sound.

First of all, the musical structure does not have to be analyzed at the level of the score, but it already becomes apparent after all mathematical operations have been applied, but before the resulting sound events are determined. These do not yet represent sound events, but indices of potential sound events (see Figure 1). As a result of the successively applied operations, generally several superimposed structures appear. By looking at the individual intermediate results, one already obtains an analytical view of the structure without additional effort.

Some operations can easily be related to known musical forms, for example a subtraction applied to $id(\mathbb{N}_0)$ corresponds to the emergence of the same sequence only time-delayed and thus to the structure of an imitation canon, e.g. $\{0, 1, 2, 3, 4, 5, 6, 7, 8, ...\} - 2$ results in $\{0, 0, 0, 1, 2, 3, 4, 5, 6, ...\}$.

As already mentioned above, values smaller than zero are always set to zero and decimal places are neglected.

The modulo division is mainly responsible for the repetitive structures that frequently occurs here, e.g. $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ...\}$ *modulo* 5 results in

$\{0, 1, 2, 3, 4, 0, 1, 2, 3, 4, ...\}$.

The division applied to $id(\mathbb{N}_0)$ results in a slowed down sequence of the same indices when successive identical indices are joined together, e.g.

$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ...\}/2$ results in $\{0, 0, 1, 1, 2, 2, 3, 3, 4, 4, ...\}$. Thus, the structures occurring during division show a certain similarity with the musical structure with the musical form of an augmentation canon.

On the whole, the actual minimal music effect results from the fact that the surgery through a newly applied operation always torments the entire picture to a not too extreme extent, instead of changing individual things in isolation.

2.4 Analysis of the Harmonic Structure

The creation of musical structures with **AOG** is constructive. There is no harmonic analysis and no correction of the harmonic interactions. This is also not necessary for **two** reasons:

The individual numbers in the sequence obtained from an arithmetic operation are used as divisors of the so-called base number. Thus, these numbers result in a certain picking of prime factors from the base number. The product of the selected prime factors – respectively the result of the division – is then interpreted as the frequency of the tone to be heard. Finally, this frequency is mapped to the tempered tone system.

The frequencies that can be generated in this way have only a limited degree of dissonance to each other. Leonhard Euler has already provided a method to measure this. He called his method "gradus suavitatis" g. It is very well suited to this task because, like the approach here, it is based on integer frequencies f, which are then broken down into their prime factors p_i to obtain their degree of dissonance g: For $f = \prod_{i=1}^{n} p_i^{k_i}$ the "gradus suavitatis" is $g = 1 + \sum_{i=1}^{n} p_i.k_i - \sum_{i=1}^{n} k_i$ [1]. (The much discussed problems in the application of the gradus to classical harmony theory will be ignored in this context.)

The degree of dissonance between two frequencies is then the gradus function for the prime factors in which the two compared frequencies differ from each other. Since the prime factors of the base number consist of relatively small prime numbers, it is immediately obvious that when comparing two frequencies that can be generated from, only relatively small degrees of dissonance are produced according to the gradus function.

In addition to this fundamental limitation of the degree of dissonance, a **second factor** that plays a role, that the harmonic event that results in **AOG** generally seems to be reasonable. It can be found in a meaningful organization not only of sound events, but also of their harmonical relationships by the algorithm.

It is anything but trivial to explain here what makes sense and what does not. Since the examples of Bach's monophonic polyphony and at the latest since the tintinabuli harmony of Arvo Pärt, it is clear that even sound events that are far apart can be related melodically or harmonically to each other if they are in the same register in the first case and even not in the second.

Since the sequences of numbers resulting from the arithmetic operations are applied as divisors of the basic number used, **AOG** does not only result in a multi-level rhythmic musical structure right from the start, but they also bring the harmonic relationships of the tones into a rhythmic order.

This principle will be illustrated in a small (a bit academic) example: The base number is $b = 2{\cdot}3{\cdot}5 = 30$. For the sequence $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ modulo 7 is applied. The result is the sequence $\{0, 1, 2, 3, 4, 5, 6, 0, 1, 2, 3, 4, 5, 6\}$.

Each sequence element is used as a divider of b. Where this is not possible, a zero remains (no tone). This results in the following frequency sequence: $\{0, 2 \cdot 3 \cdot 5, 3 \cdot 5, 2 \cdot 5, 0, 2 \cdot 3, 5, 0, 2 \cdot 3 \cdot 5, 3 \cdot 5, 2 \cdot 5, 0, 2 \cdot 3, 5\}$.

If one looks at the prime factors of the occurring tone frequencies separately, it turns out that not only the frequencies of the sound events themselves occur in rhythmic order, but also the individual prime factors of these frequencies: $\{0, 2, 0, 2, 0, 2, 0, 0, 2, 0, 2, 0, 2, 0\}$, $\{0, 3, 3, 0, 0, 3, 0, 0, 3, 3, 0, 0, 3, 0\}$,

$\{0, 5, 5, 5, 0, 0, 5, 0, 5, 5, 5, 0, 0, 5\}$.

2.5 Musical Interpretation and Sound Generation

In principle, the fact that the entire intermediate stages of the generative process on the way from the symbolic representation to the representation of the sequence of the sound events are available offers a multitude of starting points for controlling musical parameters in the field of musical interpretation. In particular, it is possible to take into account by which partial sequence of the applied operations a certain tone was produced.

However, since this work is currently less focused on this last step of musical interpretation, a rather minimalist procedure was initially applied here: Each note is assigned a sample of a percussive instrument. If it turns out that the same frequency has to be played several times simultaneously at a certain point in time, these events are just played in combination and thus form acoustically one event with a corresponding increase of volume. The whole software was implemented in Java (Processing) and for the actual sound generation a simple sequencer, which is also implemented in Java, is responsible, which allows to stream wav files (also superimposed).

3 The Concept of Transparency Considering Generative Grammars

> " *I am giving a performance in Toronto ... I call it Reunion. It is not a composition of mine, though it will include a new work of mine, 0'00" II, ...* [10]. "

John Cage represents to an extreme degree an attitude towards the work in which the maker, the composer, steps back behind the work. This attitude can be read from his late works in that arrangements of things found by chance often form the basis for a musical structure. This basic attitude has strongly influenced the art world both in the visual arts and in music, and the trend is that the composer is no longer the creator of a musical structure, but determines the setting in which the composition then happens [13]. Especially in the field of algorithmic composition there is the widespread basic attitude that the composing subject has no direct imaginative access to what the algorithm itself produces. During the discussion on [18] Sever Tipei notes that music is experimental for him when the result of the generation process is unpredictable. One

may or may not follow this paradigm, the fact is that the creation of a setting creates a certain void, which is then often filled by interaction with the (active) recipients. And it is also a fact that these people who are involved in the artistic process bring their own ideas about what music or art is. If one admits this and takes it seriously, and thus gives human interaction a higher meaning than that of a mere random generator, the question immediately arises to what extent the setting provided allows the active recipient to consciously design a (musical) performance according to his or her own ideas.

As mentioned above, in terms of Chomsky's division of generative grammars, **AOG** is one of level 3: Its application ensures that only meaningful/well-formed musical structures are created ([2], [3]) and can thus in principle also enable people with little knowledge of music theory to compose. On the other hand, this advantage is bought at the price of a certain lack of transparency with regard to the relationship between a sequence of symbols and the musical form they represent, and is thus directly opposed to the claim of being able to mentally foresee the resulting musical structures.

Can this shortcoming in **AOG** somehow be compensated by the fact that we are already well versed in dealing with arithmetic operations and infinite sequences, which together form the basis for **AOG**, due to our school education in general? So does this kind of mathematical education in **AOG** allow us to mentally understand the connection between symbolic representation and the musical form it represents?

In order to prepare an answer to this question by first gaining an approximate understanding of how a corresponding mental process can be imagined, a suitable description Husserl's will first be referred to below. It deals with the mental process of how we obtain out of an arithmetic term an idea of the set represented by it.

3.1 Husserl's "Philosophy of Arithmetic"

Husserl's "Philosophy of Arithmetic" comes from a time before he founded his phenomenological method.

The starting point for the development of mathematical concepts in this text is the set as a phenomenon directly accessible to man.

This fact alone should legitimize a deliberately phenomenological reading of this early text, as it is carried out here below. This attitude is also supported by the work of Lohmar [12], and also by the fact that Husserl again, in his later work "logical investigations" ("*Logische Untersuchungen*"), which co-founded the phenomenological method, cites the example of the mental unfolding of mathematical expressions down to the set (see below) to illustrate the difference between the instant imagination of a phenomenon ("*eigentliche Vorstellung eines Phänomens*") and a symbolically intermediated imagination ("*uneigentliche Vorstellung eines Phänomens*") [6].

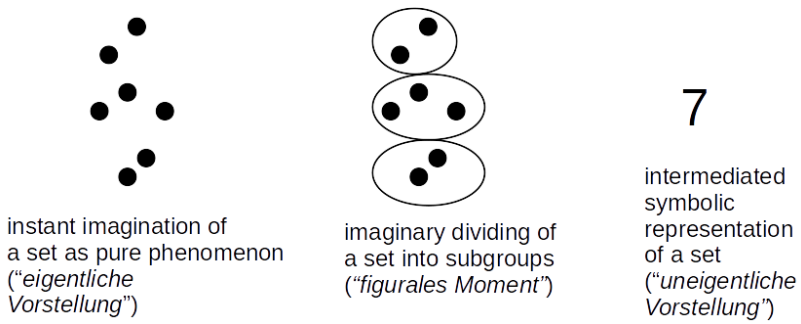


Fig. 2. How to imagine sets.

3.2 Imaginating Sets

Husserl regards sets as an elementary phenomenon. He emphasizes that an instant imagination of sets is possible [5, 201-203], but only for very small sets. And even for very small sets, we still manage by dividing them into subgroups in order to capture their extent ("*figurales Moment*") [5, 203-210] (Fig. 2).

In the course of human history, number systems have become the symbolic representation of sets and also of mechanized procedures which operate on these numbers (arithmetic), in order to merge the different sets behind them (addition), to merge several sets of the same size (multiplication), etc.

According to Husserl, the reason for this is our mental inability to perform these operations directly on the sets [5, 239-240].

3.3 The Stepwise Unfolding of Arithmetic Expressions to the Set Represented by Them

After Husserl the set is the elementary phenomenon and that the representation of numbers in the place-value system is a symbolic representation of this set, from which this set can be recovered at any time. Again, arithmetic expressions are symbolic representations, from which a certain number can be obtained unambiguously. As already mentioned above, Husserl also explains this fact at the end of the second part of his "Logical Investigations" in order to explain the representational meaning of symbols. He explains in an exemplary way (translated from German original):

> "*We make the number* $(5^3)^4$ *clear to ourselves by falling back to the definitory idea: 'Number which arises when one forms the product of* $5^3 \cdot 5^3 \cdot 5^3 \cdot 5^3$*'. If we want to make this latter idea clear again, we have to go back to the sense of* 5^3*, i.e. to the formation* $5 \cdot 5 \cdot 5$*. Going even further back, 5 can be explained by the definition chain* $5 = 4 + 1, 4 = 3 + 1, 3 = 2 + 1, 2 = 1 + 1$." [6]

.

3.4 Fulfillment Chains

In the course of the following explanations in [6], Husserl generalizes the step-by-step process of the unfolding of arithmetic expressions described here and postulates that it typically leads to an increase in the richness of content if one, starting from an imagined idea, arrives at an actual representation of a phenomenon over several unfolding steps.

An example of an imagined idea could be the memory of the name of a particular person and the actual representation of a phenomenon could then be to vividly imagine the person to me.

The area of validity of this description shall not be discussed further at this place, but only its applicability to the area of interest here. For this area it can be said without further ado: The transformation of symbolic expressions into musical structures is clearly a process in which a structure containing relatively little information is transformed into a structure with a larger amount of information (see again Figure 1). If this process is also mentally reproduced, this basically corresponds to the scheme of gradually increasing abundance described by Husserl and called "fulfillment chains" ("Erfüllungsketten" GE) by him.

The prerequisite for this information enhancement is always the availability of suitable prior knowledge: With generative grammar, I know how the algorithm works. In the example mentioned above, I remember details of the person to whom the name I came across, refers. In the following we discuss to what extent the arithmetic operations of **AOG** can be performed mentally. The necessary prior knowledge consists on the one hand in the awareness of the corresponding algorithm and on the other hand in our knowledge of arithmetic.

3.5 Phenomenological Investigation

Against the background of the eye-catching parallels of the above example to the unfolding processes described by Husserl with arithmetic expressions, the representation quoted above from Husserl's work is used, so to speak, as a blueprint for the following explanations.

In the examination of the development process described in Chapter 2.2, it is noticeable that the generation of the sequence $t \equiv 7$ can still be easily comprehended. But already here it must be said restrictively that this applies only with exclusive consideration of the first sequence members of this potentially infinite sequence. Also the following multiplication of the resulting sequence by 3 can still be imagined well. At the latest, however, when trying to apply $\equiv 5$ to the preceding result, it becomes very difficult not to lose sight of the previously obtained results.

After all: With pen and paper you can create the score from the symbol series without any problems. Only here, as with every written fixation of a score, there is still the discrepancy between writing and musical interpretation.

Even though in the development of this generative grammar great care was taken to use generally familiar structures and even though the steps in the unfolding process are completely transparent in detail, here one is still far from

being able to comprehend the unfolding process in its entirety in the mind. The system of symbols on the highest level with the rules belonging to it enables the composer to produce very complex compositions very quickly. However, the price for this is that a very multi-stage unfolding process has to be passed through in order to come down to the sound level.

Basically, all of this was to be expected, too, if one realizes that in Husserl's presentation the unfolding to a single number and finally to a set is not quite easy and that in the generative grammar introduced here we are dealing with mathematical sequences, which are sets of sets. And the latter do not even form the end point of the unfolding process here, but are followed by the transformation into a score and finally into a musical performance.

Now you can ask yourself how it is even possible to generate a relatively complex score from a few symbols. Where does that come from, what is represented in figure 1 as information growth? - Obviously this unfolding complexity has been bought with a limitation of the amount of possible compositions. Because the fact that grammars of the third order provide for the rule-compliant generation of scores at the same time states that everything in structures that cannot be obtained by applying these rules cannot be represented with the respective grammar either. And what has been said applies to any generative tool. In the present case, music arises with an affinity to (repetitive) minimal music. The musical event is shaped by the metamorphosing structurally related (tone) sequences.

Overall, the use of familiar structures as the basis of a generative grammar is a qualitative prerequisite of being able to imagine the structures unfolding from the representation of symbols, but the practical implementation fails due to the relative limitations of the human imagination.

What has been disregarded in the entire consideration so far is the possibility given today of enabling an immediate sonic implementation of symbol writing via a software in which an arbitrary change of the symbol representation instant is expressed in a corresponding sonic one. Through this feedback mechanism between generative tool and composer, an intuitive knowledge of the direct connection between symbol and sound is established over cycles of intensive use. The compactness of the symbol notation introduced here plays an important role here: it creates a good overview of the entire musical structure on an abstract level and supports the consciously executed influence on the sound event. Even further thought, over time a synaesthesia between symbolic structure, sound and emotional feeling arises, as expressed literarily in the following description of a chess game in Nabokov's "The Defence":

> " *He saw then neither the Knight's carved mane nor the glossy heads of the Pawns – but he felt quite clearly that this or that imaginary square was occupied by a definite concetrated force, so that he envisioned the movement of a piece as a discharge, a shock, a stroke of lightning – and the whole chess field quivered with tension, and over this tension he was sovereign, here gathering in and there releasing electric power.* [14]. "

4 MODULO

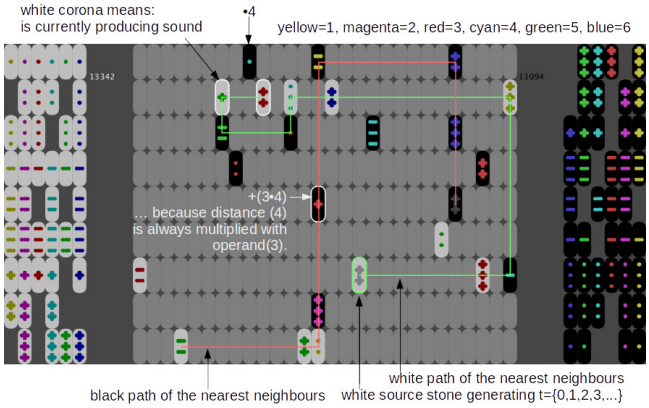


Fig. 3. View of the MODULO playing field.

But there is even another way to learn how to control the power of generative grammar:

In the computer-based board game MODULO (Fig. 3), the players behave consistent to the gaming-rules and do not care about the evolving musical structure. In this respect, MODULO represents an alternative draft to a reasonable and common use of the symbols of the grammar in which the user anticipates the musical result.

In MODULO, the game pieces are arithmetic operations. These are applied along a path of the shortest adjacent distances starting from a source tile representing $id(\mathbb{N}_0)$. Thus, such a path can be understood as a symbolic representation of a piece of music in the sense of the example given in Chapter 2.2.

Above this level is the level of the game rules for the two-person game, who alternately place tiles on the board or move them. The goal of the game is to establish an own path by skilful moves, which consists of operations and operands as mutually different as possible and at the same time to prevent the opponent from doing so. Points are awarded after each move. One way to prove that the rules of the game have been chosen in a meaningful way, as far as the resulting musical result is concerned, is to prove a positive correlation between the number of points achieved by both opponents in a game and the quality of the resulting music. In order to be able to make at least a preliminary statement about this, the game was extended by a component, in which the moves are carried out automatically, whereby from the multitude of possible moves one is always selected, which results in relatively many points. The quality of the resulting sound result can at least be seen intersubjectively in a video [8].

The moves of the opponents have a direct influence on the resulting paths and thus directly on the musical events. A move can have a metamorphosing

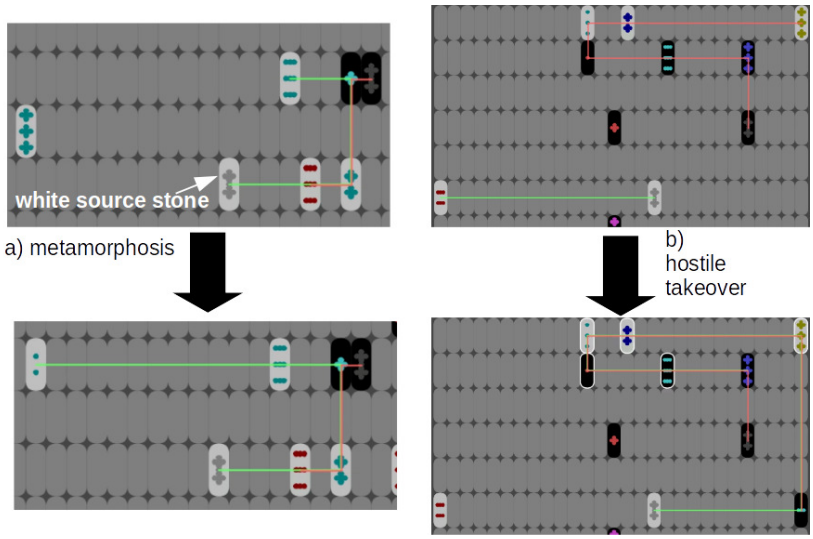


Fig. 4. a) Addition of an operation towards an existing path (metamorphosis). / b) Switching path by adding element close to source tile (hostile takeover).

character (metamorphosis, Fig. 4 a), but it can also cause drastic changes (hostile takeover Fig. 4 b).

5 Summary

It seems impossible in principle that a powerful generative grammar to be presented in a compact way is at the same time designed in such a way that the structures unfolded from it can also be imagined mentally. Theoretically this is possible, but in practice it fails because of the limitations of human imagination. At the same time, an increase in the power of the symbolic language is always linked to a restriction of the overall structures that can be generated. The fact that the symbolic representation does not correspond to the pure phenomenon, but only represents it and thus conceals it, is the reason why generative grammars are powerful tools for the composer, but can in principle not guarantee good control over the sound process, i.e. control based on knowledge. One way of actively establishing the connection between symbolic representation and sound, however, is to present both to the composer coincidentally (real-time composition tool), trusting that the composer can thus learn this connection as intuitive knowledge.

A second way is to make what makes sense measurable and then to give this measure to the composer as feedback and to trust that the composer learns at some point to intuitively maximize this measure through his actions. Such a thing takes place in a sonified, competition-driven performance, if a really meaningful connection between the rules of the game and the sound events has

been established. Thus, MODULO is integrated into a series of sonified games in which an attempt is made to establish a clear connection between the course of the game and its musical implementation [16], [4]. As a special feature in comparison to the listed examples it has to be emphasized once again that the game structure and the game rules of MODULO were obtained directly from musical considerations. Specifically, sequences of a grammar based on arithmetic operations are generated with the help of the game moves.

5.1 About Virtuososos and Sumo Robots

Unfortunately, it must be said that this work does not end with the solution of a problem, but in the best case with its sharper contouring:

The possibility to execute real-time composition either leads to trivial results if one has complete knowledge about what one is doing, or symbolically complex actions are triggered, whose non-trivial, but in the best case interesting results will never be completely transparent, especially not in real-time. In fact, however, at least the culture of classical music seems to live from the illusion that the virtuoso interpreter would react spontaneously and knowingly, for example, to the orchestra accompanying him: Through constant repetition of the same phrases in a piece, musicians learn to master a piece of music from a meta-level and can put emotional expression into these phrases, while the actual mechanical process of instrumental playing sinks into the subconscious and is thus mastered perfectly. The recipient, on the other hand, lets himself be drawn into the illusion of a spontaneous, fully conscious play in classical concerts: The enjoyment of a musical performance lies above all in experiencing the totality of technically perfect playing and apparently spontaneous emotional expression as a real fact.

While the virtuoso concert creates the illusion that the musical event unfolds directly from the moment, the illusion in automatically created compositions that are realized in real time lies in the fact that no consciously acting individual is the cause of the musical event. We only project consciousness into the machines [11]. And while in the virtuoso concert the task of the classical composer is above all to anchor the illusion of spontaneity in the structural arrangement of the composition, as a consequence of the preceding considerations the task and special challenge of the developer of real-time composition programs can be seen above all in evoking the illusion of consciously made musical decisions in the recipient.

This is where the embodiment comes into play: A box with a loudspeaker is hardly seen by the recipients as a source for conscious decisions, whereas a humanoid robot, which plays a musical instrument much more likely. This statement needs further explanation: Why do people watch competitions between sumo robots [15]? - It is the fascinating speed with which the opponents (robots) try to push each other from the battlefield. The spectators project consciously acting beings into these opponents. Such a substitution and a transcendental aspect seem to be the two basic ones at most, if not all cultural events: A lot of people come together. On a stage, something emerges that the audience would

not normally be capable of representing them. The challenge for those acting on stage is to stage this illusion as perfectly as possible.

5.2 Further Work

Based on the above, there are two possible extensions for the use of a symbolic composition based on arithmetic operations: While retaining the competition as the basic element, which allows for any kind of dramaturgy and evokes an emotional participation of the recipients, two further layers extending the previous concept would be conceivable and have already been partially implemented on a test basis:

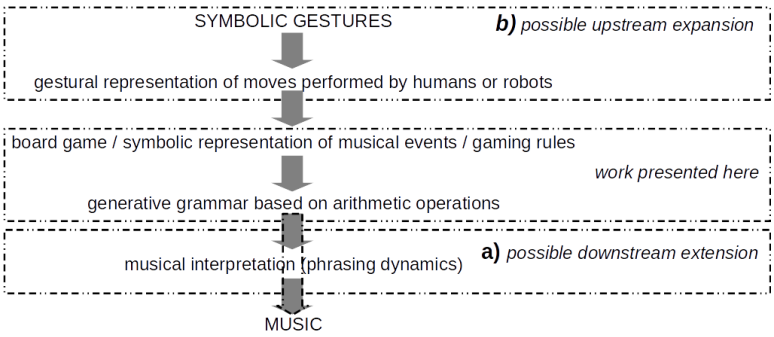


Fig. 5. Possible downstream (a) and upstream (b) expansions.

a) The previous structure can be followed by an interpretative level, which interprets the resulting musical phrases musically, i.e. complements phrasing and dynamics horizontally on the basis of the melody in the individual voices and vertically on the basis of the harmonic development. (Again, there is an increase in information due to prior knowledge, now due to known musical conventions from a certain cultural circle.)

b) An embodiment level can be inserted upstream of the previous structure, which physically implements the actions of the opposing players and thus makes them more tangible for the recipients. While in a) the previous structure can easily be supplemented (see the examples here [9]), in b) the whole structure has to be rethought: Physical movements must replace moves on a playing field. And if the result is not only to end in a mickey mousing, in which certain movements simply evoke certain sounds, but certain gestures become symbols that have a lasting influence on the sound, the symbolic interaction on the playing field must now be reworked into a symbolic, gestural interaction between two opponents. One example for an interaction of two opponents with gestural symbols is the game of rock-paper-scissors, another one can be seen in "tai chi", where interlocking movements take place between certain symbolic fighting postures [19], [20].

References

1. Busch, H.R.: Leonhard Eulers Beitrag zur Musiktheorie, p.34. Gustav Bosse, Regensburg (1970)
2. Chomsky, N.: Three Models for Description of Language. In: Information Theory, IRE Transactions, 2(3),113-124 Retrieved April 4, 2019, from `https://chomsky.info/wp-content/uploads/195609-.pdf`, (1956)
3. Chomsky, N.: On Certain Formal Properties of Grammars. In: Information and Control, 2, 137-167, Retrieved April 4, 2019, from `http://twiki.di.uniroma1.it/pub/LC/WebHome/chomsky1959.pdf` (1959)
4. Hamilton R.: Musical Sonification of Avatar Physiologies, Virtual Flight and Gesture. In: Sound, Music, and Motion (CMMR 2013), pp. 517–532, Springer, Heidelberg (2014)
5. Husserl, E.: Philosophie der Arithmetik, Martinus Nijhoff, The Hague (1970)
6. Husserl, E.: Stufenreihen mittelbarer Erfüllungen. Mittelbare Vorstellungen. In: Logische Untersuchungen, second part, pp. 601–602. Felix Meiner, Hamburg (2009)
7. Kramann, G.: "$\equiv\ 7 \cdot 5 \equiv 3 \cdot 5$" – a simple composition as an example, from `http://www.kramann.info/cmmr2019a` (2019)
8. Kramann, G.: M O D U L O `http://www.kramann.info/cmmr2019b` (2019)
9. Kramann, G.: Possible downstream and upstream expansions `http://www.kramann.info/cmmr2019c` (2019)
10. Kuhn, L.D. (ed.): The selected letters of John Cage, p.382. Wesleyan University Press, Middletown (2016)
11. Leidlmair, K.: Künstliche Intelligenz und Heidegger – Über den Zwiespalt von Natur und Geist, Wilhelm Fink, Paderborn (1999)
12. Husserls Phänomenologie als Philosophie der Mathematik. Doctoral dissertation, Faculty of Philosophy, University of Cologne, Cologne (1987)
13. Motte-Haber, H. de la: Selbständigkeit als Prinzip künstlerischer Settings. In: Neue Zeitschrift, 6, 52-56 (2018)
14. Nabokov, V.: The Defence, pp. 72-73, Granada Publishing Limited, London (1971)
15. McGregor, R.: Robot Sumo. Retrieved April 4, 2019, from `https://www.youtube.com/watch?v=QCqxOzKNFks` (2017)
16. Sinclair S., Cahen R., Tanant J., Gena P.: New Atlantis: Audio Experimentation in a Shared Online World. In: Bridging People and Sound (CMMR 2016), pp. 229–246, Springer, Heidelberg (2017)
17. Supper, M.: A Few Remarks on Algorithmic Compositions. In Computer Music Journal, 25(1), 48–53 (2001)
18. Tipei, S.: Emerging Composition: Being and Becoming – An Experiment in Progress. In Proceedings of the Sound and Music Computing Conference 2016, Hamburg. Retrieved May 22, 2017, from `http://smcnetwork.org/system/files/SMC2016_submission_73.pdf` (2016)
19. Wama T., Higuchi M., Sakamoto H., Nakatsu R. Realization of Tai-Chi Motion Using a Humanoid Robot. In: Jacquart R. (eds) Building the Information Society. IFIP International Federation for Information Processing, vol 156. Springer, Boston, MA. Retrieved April 4, 2019, from `https://link.springer.com/content/pdf/10.1007%2F978-1-4020-8157-6_9.pdf` (2004)
20. Zhuang, H.:The Mind Inside Yang Style Tai Chi: Lao Liu Lu 22-Posture Short Form. YMAA Publication Center, Wolfeboro (2016)

A phenomenological approach to investigate the pre-reflexive contents of consciousness during sound production

Degrandi M., Mougin G., Bordonné T., Aramaki M., Ystad S., Kronland-Martinet R., and Vion-Dury J.

Aix Marseille Univ , CNRS, PRISM , Marseille, France
marie-monique.degrandi@ap-hm.fr

Abstract. This article describes a listening experiment based on elicitation interviews that aims at describing the conscious experience of a subject submitted to a perceptual stimulation. As opposed to traditional listening experiments in which subjects are generally influenced by closed or suggestive questions and limited to predefined, forced choices, elicitation interviews make it possible to get deeper insight into the listener's perception, in particular to the pre-reflexive content of the conscious experiences. Inspired by previous elicitation interviews during which subjects passively listened to sounds, this experience is based on an active task during which the subjects were asked to reproduce a sound with a stylus on a graphic tablet that controlled a synthesis model. The reproduction was followed by an elicitation interview. The trace of the graphic gesture as well as the answers recorded during the interview were then analyzed. Results revealed that the subjects varied their focus towards both the evoked sound source, and intrinsic sound properties and also described their sensations induced by the experience.

Keywords: elicitation interview, auditory perception, sound synthesis, graphic gestures

1 Introduction

When preforming perceptual evaluations of sounds, it is important to be aware of the fact that listeners may focus on different aspects. Gaver [5] distinguished everyday listening from analytical listening. In the case of everyday listening of a simple source, the listener pays attention to the sound producing object, such as its size [9] and the material of which it is composed [7], [1]. In the case of more complex situations reflecting for instance interactions between sound sources, the listener perceives properties related to the event as a whole. Warren and Verbrugge [25] showed that objects that bounce and break can be distinguished by listeners with a high degree of accuracy, while Repp [15] revealed that subjects were able to recognize their own recorded clapping and the hand position from recordings when someone else is clapping. More recently, Thoret et al. [22, 21] showed that subjects were able to recognize biological motions and certain

shapes from friction sounds produced when a person is drawing on a paper.
To favor analytical listening where the listeners focus on intrinsic sound properties linked, for instance, to loudness, pitch, and timbre other approaches have been used. Merer [12] used acousmatic sounds for which the source could not be easily recognized to reveal sound structures responsible for the evocation of movement categories. Other approaches such as sensory analysis during which a group of subjects identify sensory descriptors such as onomatipoeias have been used, for instance to characterize the formantic transition from "ON" to "AN" that characterizes sounds from car motors [16, 19].
Other approaches, such as vocal imitations, that do not specifically focus on everyday or analytical listening have been used to extract relevant features of kitchen sounds [10], and more recently to reveal invariant structures responsible for the evocation of movements and materials [3]. Psycholinguistic analyses have been used to characterize sounds from musical instruments through spontaneous verbalizations. One such study that investigated violinists' preference judgements during a playing task, led to a model that linked auditory and haptic sensations to the timbre, quality, and playability of the instrument [18, 17].
Sound perception is a conscious experience that can be described not only in so-called "third person" protocols (from the point of view of the experimenter within a given paradigm, e.g. a psycho-physical paradigm), but also by protocols aiming at describing the experience from the subjects' perspective (subjective methods) mainly based on the Husserlian phenomenology. Most of the time, spontaneous descriptions of experiences and cognitive processes are poor [14] because the experience does not guarantee immediate access to its background contents [23]. Several kinds of information usually remain undisclosed, masked or "pre-reflexive" as they are called in phenomenological language [14]. Various methods allow to accurately describe the conscious experience in its reflexive and mostly pre-reflexive part. Among them, the elicitation interview (EI) [24, 11] is a disciplined introspection method conceptually based both on neurolinguistic programming (NLP) and Husserlian phenomenology [8]. EI makes it possible to return to the non-reflexive part of the conscious experience of a subject, hereby limiting influences from closed or suggestive questions.
Whereas the qualitative research methods used in sociology, such as Glaser and Strauss' anchored theory (see [17]) or the "repertory grid" method use textual corpora of reflexive descriptions of experiences to extract emerging themes and their variations, EI is essentially interested in the non-reflexive component of the experience. For this reason, whereas in the qualitative methods, the subjects use their autobiographical memory, in the EI, the subjects must relive their experience and activate their "integral memory", in particular corporeal.
We previously described pre-reflexive conscious experiences in passive listening of sounds [13]. In the current work we analyze pre-reflexive content of conscious experiences in an active task consisting in reproducing a sound by drawing on a graphic tablet.

2 Material and methods

In this section we describe the interactive device used by the participants, the experimental protocol and the elicitation interview.

2.1 Equipment: The "tablet-synthesizer" device.

Sound synthesis is a powerful tool to create any kind of sounds that either imitate real or virtual situations. Current synthesis models enable high quality re-synthesis of natural sounds that can be generated in real-time. One challenging aspect linked to sound synthesis is the control of the synthesis parameters that is not always intuitive. To meet this challenging control issue, we have developed a synthesizer based on perceptual features linked to the evocation of actions and objects [2, 1]. This device is based on the ecological approach to perception proposed by Gibson [6] which considers that actions and objects are recognized through invariant structures. The sound synthesizer makes it possible to create sounds from verbal labels that describe the action (e.g. hitting, scraping, rolling) and the object (e.g. material, size, shape) associated with the sound. Any combination between actions and objects can hereby be simulated, such as scratching a small metallic bell or hitting a big wooden bar [4]. Unrealistic situations can also be simulated this way, such as rubbing the wind or scratching a wave.
In the present study we decided to use a sound texture that evoked a movement in water, since the timbre of liquid sounds vary strongly with the dynamic action. To create the reference sound that the subjects were asked to reproduce, the synthesized sound was combined with an elliptic movement recorded by the experimenter who drew on a WACOM INTUOS PRO graphic tablet. The experimenter freely chose the eccentricity and the orientation of the ellipse that he/she was asked to draw ten times. To induce a periodic movement, we used a 60 bpm metronome while the experimenter was drawing to help him/her maintain a regular speed. Among the ten repetitions, the three most regular ellipses were selected. The position of the stylus was recorded by a Max/MSP interface at a sampling rate of 129 Hz. We then derived the position to get the velocity profile.

2.2 Experimental protocol

The subjects were first asked to listen to the reference sound which nature and origin they ignored. They were then asked to reproduce this reference sound on the WACOM INTUOS PRO graphic tablet with the gesture that best imitated the reference sound. The subjects produced the sound in real time while they performed the gesture on the graphic tablet.

Participants Ten subjects, 7 women and 3 men (aged from 26 to 70 years) were included in this experiment. Five subjects were experienced musicians practicing

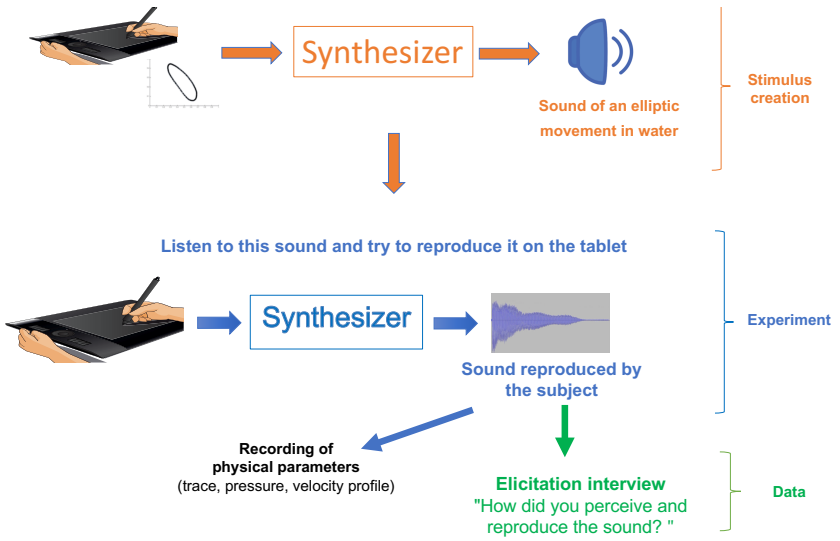


Fig. 1. Experimental protocol

an instrument on a regular basis and the the remaining 5 participants were not musicians. The ten subjects were right handed. Subjects did not have any hearing or neurological problems, such as memory-related problems or attention difficulties. The interview was conducted by one of the three doctors involved in the study: MD, GM, JVD. An audiogram was performed for each subject before the beginning of the experiment to make sure that none of the subjects had hearing impairments.

The elicitation interview In a second step (just after the reproduction of the sound), the subjects were asked to review their experience while listening to and reproducing the sound by means of an elicitation interview, by answering the question "how did you perceive and reproduce the sound?". The EI was conducted by three experienced researchers in phenomenology and EI.
The EI requires a certain number of methodological specificities:

a) The first key of the interview is to lead the subjects to describe their experience, that is to tell what they experienced and not what they thought, believed or imagined to have been their experience [14].

b) The interviewer should lead the subjects to discuss their past experiences by helping them to find the sensory and emotional dimensions.

c) The interview consists in helping the subjects redirect their attention from the content of their experience (the "what"), to its diachronic and synchronic

structure oriented towards the experiential (non-causal) “how”. The diachronic structure of the experience corresponds to the stages of its deployment over time. The synchronic structure of the experiment corresponds to the configuration at a given moment of the sensory registers used, the type of mobilized attention... etc. The aim is to make the subjects relive their experience rather than to remember it.

d) To collect such a description, the interviewer's questions should be “empty of content”, non-inductive and “point” to the structure of the experiment without providing any content. Questions are, for example: “From what did you start? What did you feel ? How did it appear to you? ”, etc. This mode of questioning emphasizes the “how” of the conscious experience and excludes the “why”.

e) The structure of an interview is iterative while guiding the attention of the subject towards a diachronic or synchronic mesh which progressively becomes more detailed each time. The average duration of an interview is about an hour to describe a few seconds of experience (as Stern puts it, “there is a world in a grain of sand” [20]). The interviewer must remain totally neutral. A good harmonization of affects (motor and prosodic affective tuning [20]) is a critical condition for the quality of the interview.

Data collection and analysis. All the EI were recorded, with the subjects' agreement. The physical data (pen movement, speed, pressure etc ...) were collected from the computer connected to the graphic tablet. The records of EI were entirely transcribed. The analysis of verbatim was carried out to extract the descriptive categories (saliencies) from each interview. The choice of descriptive categories for each interview was validated by 7 people in an inter-judge session.

3 Results

The physical data from the tablet were analysed together with the EI. Only the data from the EI, as well as the drawings recorded on the tablet are presented in Tables 1 and 2.

Types of sound listening. The EI enabled to collect the synchronic and diachronic structure from the listening experience of each subject. These data respond to both the ”what” of their experience but also to ”how”, to the proper way of perceiving and reproducing this sound. They give a fine and precise description of an experience that lasted for a few seconds by allowing an awareness of the different processes.
Four descriptive categories (attractors) which are common to all the subjects can be identified. These categories are related to the way the subjects hear a sound while they prepare its reproduction: 1) the direction of listening, 2) the

Table 1. Three types of listening experiences

Types of listening experiences						
Listening focus (LT)	**Main LF used by the subject**	**Nomber of subjects using this LF**	**Main sensory modalities involved**	**Attentional disposition**	**Sound-auditor positon**	**Moment of appearance**
Origin of sound	2, 4 and 5	8/10	Scenes (sea, beach…) perceived by the auditory and visual modalities	Directed attention towards the source. Active search remembrance, familiar scenes evoking the source.	Location of the subject in relation to the scene.	Appears spontaneously first while listening
Acoustic characteristics of sound	1, 7, 8 and 10	8/10	Timbre, intensity, rhythm, height perceived by the auditory modality, but may be associated with other modalities (rhythm with kinesthetic sensitivity)	Attention directed towards the different parts of the sound Active position of the subject in relation to the sound.	Accurate location of the sound, external to the subject....	Appears when subjects focus on the task of reproduction
Effect of the sound	3, 6 and 9	8/10	Dynamics of the sound mainly perceived by kinesthetic sensitivity	Attention less focused, more global Position of the subject rather passive compared to the sound	Blurred boundary between body space and sound Effect of sound throughout the body.	Particular listening modality, generally not described spontaneously rather evoked at the end of EDE.

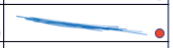
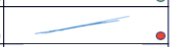
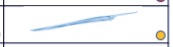
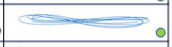
sensory listening modalities, 3) the attentional disposition 4) the reproduction strategies. The first three descriptive categories which correspond to three types of listening are in line with categories identified in our previous work [13]. The fourth is specific to this study. Each of the 3 types of listening can be analyzed from a) the main sensory modalities used, b) the attentional disposition of the subject, c) the position of the subject with respect to the sound and d) the moment this type of listening occurs. Each subject has a preferred type of listening (in this experiment), but this does not mean that he or she does not use other types of listening in a less marked way. This part of the analysis is presented in Table 1.

The first type of listening is turned to the source of the sound and involves attention directed to the origin of the sound with an active search for familiar scenes associated with the source. In this type of listening the imagination is very active. The subject is thus projected into an imaginary scene evoked by the sound heard which is integrated into the scene, and a given context in the visual modality. This listening structure appears spontaneously and early in the diachronic description of the experience. This type of listening, characterized as everyday listening by Gaver [5], represents the main listening mode of three subjects but is, for 8 out of 10 subjects, associated with the other types of listening.

The second type of listening, characterized as analytic listening by Gaver, is directed to the characteristics of the sound. This way of perceiving sounds appears when subjects focus on the reproduction task. This time the sound is brought back to its different components (rhythm, pitch, timbre, intensity), and the subjects focus on the sound itself and not on the causality. This is the main listening type for four subjects, but 8 out of 10 subjects used it in the experiment.

The third level of listening is a particular listening modality that is usually not spontaneously described in our daily lives and rather evoked at the end of the diachronic description of the listening experience. This is a way of listening that focuses on the effect of the sound, specifically the dynamics, the movement it induces relative to the whole body. It is an "internal" or "embodied" listening modality in which the boundaries between the sound and the corporal space become porous. Subjects adopt a more passive position related to the sound, in a way they are "impregnated by the sound". This is the main listening modality of three out of ten subjects, but 8 out of 10 subjects used it in the experiment. Finally, we did not find any difference between musicians and non-musicians with respect to the type of listening.

Table 2. Reproduction task and type of listening. The colored circles indicate the coherency between the representation of the sound and the imaginary content or the reproduction gesture (green = good, orange = medium, red = poor)

Listening focus (LF)	Subjects	Representation of sound	Imaginary content	Recorded trace (movement)
Origin of sound	2	Wave	Wave	
	4	Wave with bubbles	Wave	
	5	Waves with bubbles	Wave	
Acoustic characteristics of sound	1	Something perfectly rounded	Ellipse	
	7	Dynamics of the sound	Sinusoids	
	8	Rhythm	Rhythm	
	10	Wave	Wave	
Effect of the sound	3	Oscillation, oval shape	Hourglass	
	6	Dynamics of the sound	Kind of Ellipse	
	9	Pulsations of the sound	Kind of spindle	

Reproduction strategies. An original result of this study is that the representation of the sound and the imaginary content of the drawing gesture to perform depends on the major mode of listening for each subject (Table 2). Subjects with a predominant listening based on the origin of the sound (i.e. everyday listening) imagined waves. Subjects with a predominant listening based on the acoustic characteristics of the sound (i.e. analytic listening) rather considered the physical parameters with a coherent imaginary content with these parameters. Subjects with a predominant listening based on the sound effect rather felt oscillations and movements and evoked the elliptic shape in their imaginary content. Surprisingly, the actual realization of the trace is not closely coherent with the imaginary content of the gesture and it seems that it does not depend

on the preferential manner in which the sound is listened to. We did not find any relation between the type of listening and the age or gender of the subjects or between musicians and non-musicians.

4 Discussion and conclusion

The phenomenological analysis of the pre-reflexive contents of the consciousness in a reproduction task of a sound using a sound-based graphic tablet makes it possible to confirm the main types of listening previously described by Gaver [5] or Petitmengin et al. [13]. The fact of having a reproduction task to be accomplished modifies, with respect to an isolated passive listening, the diachronic and synchronic content of this experience (the moment of appearance of the experiential content, in particular).

In this preliminary work involving a small population of subjects, we did not find differences in listening and sound reproduction based on age, gender, or musical experience. It would be interesting to increase the number of subjects to assess whether differences appear according to these factors. However, we can not perform EI on large populations because of the considerable time required for data processing. We (GM, JVD) are currently testing faster and more efficient data processing methods to increase the number of subjects involved in this type of study. When comparing our current and previous studies [13], several differences must be reported. The initial study focused on describing the modalities of listening to the sound, as such, and without any task required at the end of the listening. The study aimed to highlight the descriptive categories of the non-reflexive part of the sound listening experiences and to define the general structure of such an experience. For this reason, various sounds were used (sounds from nature, sounds from everyday life, abstract sounds). Some individual differences linked to the way subjects listened to sounds were observed, but the constitution of subgroups of subjects did not appear. In our current study, only one sound is proposed with an associated reproduction task. If the same types of non-reflexive experiences can be observed, the task to be done changes the type of intentionality [8] and attentional focus. The task to be performed involves motor strategies, whereas in passive listening such strategies are absent. Moreover, the tablet-synthesizer device constrains the motor strategy and probably the associated imaginary processes.

This study also made it possible to highlight the fact that even if each subject possesses a preferential type (focus) of listening, other types of listening are also mobilized to find the resources for carrying out the reproduction task. This entanglement of available perceptual dispositions opens a new field of research on the co-presence of pre-reflexive complex processes involved. Another unexpected result is that, on the one hand, since we find a correct coherence between the preferential type of listening and both the representation of the sound in the consciousness and the imaginary content attached to the act of reproduction, on the other hand, it clearly appears in Table 2 that the actual traces recorded on the tablet no longer display the same coherences and do not appear as closely corre-

lated to the imaginary content attached to the act of reproduction. This relative dissociation suggests that in the entire audio-motor loop, cognitive and motor processes generating the drawings are not entirely constrained by the imaginary processes associated with the sound. Two hypotheses could explain this relative dissociation: 1) the audio-motor loop would have a relative autonomy compared to the construction of the motor act of reproduction and would not modify the motor control. The imaginary content would then be an epiphenomenon more or less independent but generated by the sound heard, 2) the imaginary content aroused by the sound would modulate more or less the driving act of reproduction, according to the personality of the subject, his/her interests and the context of the experimentation. This exploratory and multidisciplinary work seems to provide an early proof of concept of the use of introspective methods in acoustics and audition in order to refine synthesis models and sound control towards an approach more and more turned towards the human experience.

Acknowledgments. We thank the members of the "Atelier de Phénoménologie Expérientielle" (Marseille) for their participation to this study. This work is partly supported by the French National Research Agency and is part of the "Sonimove" project (Grant No. ANR-14-CE24-0018).).

References

1. Aramaki, M., Besson, M., Kronland-Martinet, R., Ystad, S.: Controlling the perceived material in an impact sound synthesizer. IEEE Transactions on Audio, Speech, and Language Processing 19(2), 301–314 (2011)
2. Aramaki, M., Gondre, C., Kronland-Martinet, R., Voinier, T., Ystad, S.: Imagine the sounds : an intuitive control of an impact sound synthesizer. In: Ystad, Aramaki, Kronland-Martinet, Jensen (eds.) Auditory Display, Lecture Notes in Computer Science, vol. 5954, pp. 408–421. Springer-Verlag Berlin Heidelberg (2010)
3. Bordonné, T., Dias-Alves, M., Aramaki, M., Ystad, S., Kronland-Martinet, R.: Assessing sound perception through vocal imitations of sounds that evoke movements and materials. In: Aramaki, Davies, Kronland-Martinet, Ystad (eds.) Music Technology with Swing, Lecture Notes in Computer Science, vol. 11265, pp. 402–412. Springer Nature Switzerland (2018)
4. Conan, S., Thoret, E., Aramaki, M., Derrien, O., Gondre, C., Ystad, S., Kronland-Martinet, R.: An intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling. Computer Music Journal 38(4), 24–37 (2014)
5. Gaver, W.W.: What in the world do we hear? an ecological approach to auditory event perception. Ecological Psychology 5(1), 1–29 (1993)
6. Gibson, J.J.: The ecological approach to visual perception: classic edition. Psychology Press (2014)
7. Giordano, B.L., McAdams, S.: Material identification of real impact sounds: Effects of size variation in steel, wood, and plexiglass plates. Journal of the Acoustical Society of America 119(2), 1171–1181 (2006)
8. Husserl, E.: Ides directrices pour une phénoménologie. Gallimard (1985)
9. Lakatos, S., McAdams, S., Chaigne, A.: The representation of auditory source characteristics : simple geometric form. Perception and Psychophysics 59, 1180–1190 (1997)

10. Lemaitre, G., Dessein, A., Susini, P., Aura, K.: Vocal imitations and the identification of sound events. Ecological Psychology 4(23), 267–307 (2011)
11. Maurel, M.: The explicitation interview: example and applications. Journal of Consciousness Studies 16, 20–57 (2009)
12. Merer, A., Aramaki, M., Ystad, S., Kronland-Martinet, R.: Perceptual characterization of motion evoked by sounds for synthesis control purposes. ACM Trans. Appl. Percept. 10(1), 1–24 (2013)
13. Petitmengin, C., Bitbol, M., Nissou, J., Pachoud, B., Curallucci, H., Cermolaccce, M., Vion-Dury, J.: Listening from within. Journal of Consciousness Studies 16, 252–284 (2009)
14. Petitmengin, C., Bitbol, M., Ollagnier-Beldame, M.: Vers une science de l'expérience vécue. Intellectica - Rev Assoc Pour Rech Sur Sci Cogn ARCo. 64, 53–76 (2015)
15. Repp, B.H.: The sound of two hands clapping: An exploratory study. The Journal of the Acoustical Society of America 81(4), 1100–1109 (1987)
16. Roussarie, V., Richard, F., Bezat, M.C.: Validation of auditory attributes using analysis synthesis method. In: Congrés Francais d'Acoustique/DAGA. Strasbourg (2004)
17. Saitis, C., Fritz, C., Scavone, G., Guastavino, C., Dubois, D.: A psycholinguistic analysis of preference verbal descriptors by experienced musicians. Journal of the Acoustical Society of America 141(4), 2746–2757 (2017)
18. Saitis, C., Giordano, B., Fritz, C., Scavone, G.: Perceptual evaluation of violins. a quantitative analysis of preference judgments by experienced players. Journal of the Acoustical Society of America 132(6), 4002–4012 (2012)
19. Sciabica, J., Olivero, A., Roussarie, V., Ystad, S., Kronland-Martinet, R.: Dissimilarity test modelling by time-frequency representation applied to engine sound. In: Audio Engineering Society Conference: 45th International Conference: Applications of Time- Frequency Processing in Audio (2012)
20. Stern, D.: The Present Moment in Psychotherapy and Everyday Life. W. W. Norton & Company (2010)
21. Thoret, E., Aramaki, M., Bringoux, L., Ystad, S., Kronland-Martinet, R.: Seeing circles and drawing ellipses: when sound biases reproduction of visual motion. PLoS ONE 11(4) (2016)
22. Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.L., Ystad, S.: From sound to shape: auditory perception of drawing movements. Journal of Experimental Psychology: Human Perception and Performance 40(3), 983 (2014)
23. Vermersch, P.: Conscience directe et conscience rflchie. Intellectica 31, 269–311 (2000)
24. Vermersch, P.: Describing the practice of introsopection. Journal of Consciousness Studies 16, 20–57 (2009)
25. Warren, W.H., Verbrugge, R.R.: Auditory perception of breaking and bouncing events: a case study in ecological acoustics. Journal of Experimental Psychology : Human Perception and Performance 10(5), 704–712 (1984)

Mobile Music With the Faust Programming Language

Romain Michon,[1,2] Yann Orlarey,[1] Stéphane Letz[1], Dominique Fober[1] and Catinca Dumitrascu[1]

[1] GRAME-CNCM, Lyon (France)
[2] CCRMA, Stanford University, Stanford (USA)
michon@grame.fr

Abstract. The FAUST programming language has been playing a role in the mobile music landscape for the past ten years. Multiple tools to facilitate the development of musical smartphone applications for live performance such as `faust2ios`, `faust2android`, `faust2api`, and `faust2smartkeyb` have been implemented and used in the context of a wide range of large scale musical projects. Similarly, various digital musical instruments leveraging these tools and based on the concept of augmenting mobile devices have been created. This paper gives an overview of the work done on these topics and provide directions for future developments.

Keywords: FAUST, Mobile Music, Digital Lutherie

1 Introduction

The field of mobile music has been active for the past fifteen years [1]. It started with early experiments on programmable smartphones around 2004 [2, 3] but it really took off in 2007 when the iPhone was released and smartphones started to spread out to quickly become a standard [4]. The FAUST[3] project [5] through its core developer team at GRAME-CNCM[4] involved itself in this action in 2010 with initial experiments on running FAUST programs on iOS devices. Since then, a panoply of tools to generate standalone smartphone applications and audio engines for different mobile platforms (i.e., Android and iOS) have been developed and used as part of a wide range of musical and pedagogical projects.

In this paper, we give an overview of the work that has been done around mobile music in the context of the FAUST programming language. We present `faust2ios`, `faust2android`, `faust2api`, and `faust2smartkeyb` which are tools that can be used to create musical mobile apps at a high level using FAUST. Work carried out on the idea of augmenting mobile devices with passive and active elements to turn them into specific musical instruments is described. An

[3] `https://faust.grame.fr` (All URLs presented in this paper were verified on May 2, 2019.)
[4] `http://www.grame.fr`

overview of various musical projects such as *SmartFaust*, *SmartMômes*, and *Geek-Bagatelles* is presented. Finally, we talk about current developments and future directions for this type of work.

2 `faust2ios`

Pushed by the interest around mobile music in the early 2010s (see §7), we worked at GRAME-CNCM on a tool to convert FAUST programs into ready-to-use iOS applications: `faust2ios`. As any other FAUST "architecture,"[5] the user interface of such apps is based on the UI description provided in the FAUST code, and is therefore typically made out of sliders, knobs, buttons, groups, etc.

Figure 1 presents a screenshot of sfCapture,[6] an app made with `faust2ios` as part of the *SmartFaust* project (see §7.1).

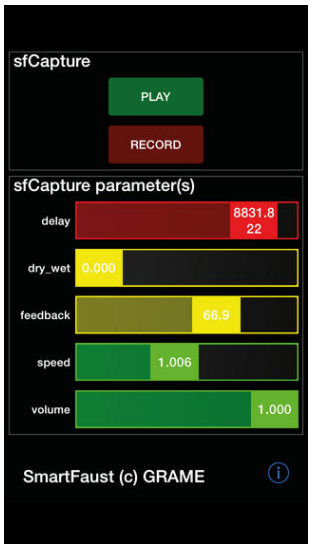


Fig. 1. Screen-shot of sfCapture, an App Made with `faust2ios`.

`faust2ios` works as a command line tool taking a FAUST program as its main argument and producing in return either a ready-to-install iOS app or the Xcode project corresponding to this app. For example, running the following command in a terminal:

```
faust2ios myFaustProgram.dsp
```

[5] Architectures in the FAUST vocabulary refer to wrappers allowing to turn a FAUST program into a specific object such as standalone desktop program, an audio plug-in, a smartphone app, an audio engine for a specific platform, etc.

[6] `https://itunes.apple.com/us/app/sfcapture/id799532659?mt=8`

will produce an iOS app corresponding to the FAUST program implemented in `myFaustProgram.dsp`.

Various features can be added to the generated app such as MIDI, OSC and polyphony support simply by using specific flags (options) when running `faust2ios`. Regular FAUST options are also available to generate parallelized DSP[7] code, change sample resolution, etc. Any parameter of a FAUST program can be assigned to a specific axis of a built-in motion sensor (i.e., accelerometer, gyroscope, etc.) of the smartphone simply by using metadata. Complex non-linear mappings can be implemented using this mechanism.[8]

Implementing `faust2ios` was relatively straightforward since the FAUST compiler can generate C++ code and that iOS applications can be implemented in Objective-C which allows for the direct use of C++.

3 `faust2android`

Motivated by the success of `faust2ios` (see §2) among composers and developers at GRAME-CNCM, we started the development of a similar system for the Android platform in 2013 [6]. This proved to be way more challenging than we anticipated, mostly because Android was never designed with real-time audio applications in mind. First, the fact that JAVA is used as the preferred programming language to develop Android apps was problematic since it doesn't perform well in the context of real-time DSP. Hence, the audio portion of the app must be implemented in C++ and the higher level elements in JAVA. This implies the use of wrappers between these two languages which is not straightforward to implement. Another issue with Android was that despite the use of low-level native code for the DSP portion of the app, audio latency used to be dreadful around 2013 (greater than 200ms), discarding any potential use in a musical context.

Despite these difficulties, the first version of `faust2android` was released in the first quarter of 2013 [6]. It had similar features than `faust2ios` (see §2) and worked in a very similar way as a command line tool. Figure 2 presents a screenshot of an app generated with `faust2android`.

As time passed and the market for real-time audio applications on smartphone grew up, Google slowly addressed the audio latency issue of Android and acceptable performances matching that of the iOS platform (less than 20ms) were achieved by 2016. Additionally, Google released in 2017 a new C++ API for real-time audio on Android which significantly simplified the design of apps involving this kind of element.[9]

On the `faust2android` front, various new features were added to replace the standard FAUST user interface of Android apps by advanced interfaces more usable in a musical context such as piano keyboards, X/Y controllers, etc. [7] These opened the path to `faust2smartkeyb` which is presented in §5.

[7] *Digital Signal Processing*

[8] `https://faust.grame.fr/doc/manual#sensors-control-metadatas`

[9] This new API is currently not used by `faust2android`, which predates its release.

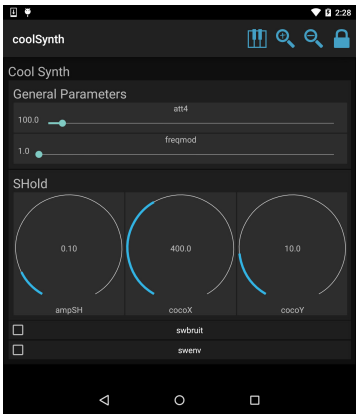


Fig. 2. Example of Interface Generated by `faust2android` Containing Groups, Sliders, Knobs and Checkboxes.

4 `faust2api`

With `faust2ios` (see §2) and `faust2android` (see §3) appeared the need for a generic system to generate audio engines with a high-level API similar across languages (i.e., JAVA, C++, etc.) using FAUST: `faust2api` [8]. The main goal of this tool was to offer iOS and Android developers with little background in audio DSP a simple way to generate ready-to-use engines for sound synthesis and processing.

`faust2api` is a command line tool working in a similar way than `faust2ios` and `faust2android`. It takes a FAUST program as its main argument and accept more or less the same options than `faust2ios` and `faust2android`. The format of the generated engines varies between platforms but the same API can be used to configure and control it.

`faust2api` was released in 2017 and was used as the basis for `faust2smartkeyb` (see §5). `faust2ios` and `faust2android` were simplified by using `faust2api` to carry out real-time audio DSP tasks. Because of its large success among developers, the concept of `faust2api` was spread to most of FAUST's targets and it can now be used to generate audio engines for desktop applications, plug-ins, etc.

5 `faust2smartkeyb`

With the latest developments of `faust2android` (see §3), we started exploring the idea of replacing the standard FAUST user interface made out of sliders, buttons, groups, etc. with more advanced interfaces, better adapted to a use in a live music performance context and to touch-screens. We extended this idea with SMARTKEYBOARD which is a highly configurable keyboards matrix where keys can be seen both as discrete buttons and continuous X/Y controllers. For

example, a keyboard matrix of size 1x1 (a single keyboard with a single key) will fill up the screen which can then be used as a multi-touch X/Y controller.

This type of interface is available as part of the `faust2smartkeyb` command line tool [9] which allows us to turn a FAUST program into an iOS or an Android app with a SMARTKEYBOARD interface. The interface can be configured directly from the FAUST code using a metadata. For example, the following program:

```
declare interface "SmartKeyboard{
  'Number of Keyboards':'2'
}";
import("stdfaust.lib");
f = nentry("freq",200,40,2000,0.01);
g = nentry("gain",1,0,1,0.01);
t = button("gate");
envelope = t*g : si.smoo;
process = os.sawtooth(f)*envelope <: _,_;
```

implements a synthesizer based on a sawtooth wave oscillator and a simple exponential envelope controlled by two parallel piano keyboards on the touch-screen (see Figure 3). Connection between the interface and the DSP part is carried out by the use of standard parameter names. Hence, `freq` is automatically associated to the pitch on the keyboard, `gain` to velocity, and `gate` to note-on/off events.[10]

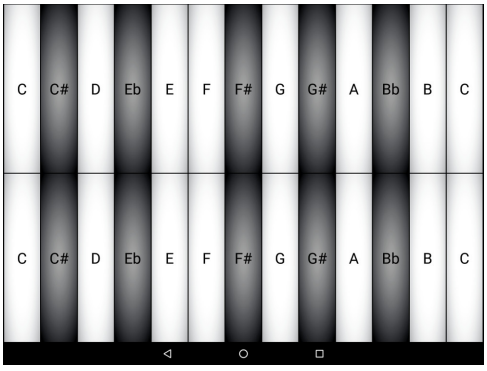


Fig. 3. Simple SMARTKEYBOARD Interface.

Complex behaviors can be implemented to handle polyphony, monophony (e.g., voice stealing, priority to upper or lower keys, etc.), and continuous pitch control (e.g., quantization, "pitch rounding" to be in tune and allow for vibrato and glissandi to be performed at the same time, etc.).

[10] `https://faust.grame.fr/doc/manual#standard-polyphony-parameters`

In the following example, a completely different app is implemented where a single key on a single keyboard is used to control a simple synthesizer producing a constant sound (no key on/off):

```
declare interface "SmartKeyboard{
  'Number of Keyboards':'1',
  'Max Keyboard Polyphony':'0',
  'Keyboard 0 - Number of Keys':'1',
  'Keyboard 0 - Send Freq':'0',
  'Keyboard 0 - Static Mode':'1',
  'Keyboard 0 - Piano Keyboard':'0',
  'Keyboard 0 - Send Numbered X':'1',
  'Keyboard 0 - Send Numbered Y':'1'
}";
import("stdfaust.lib");
//////// parameters ////////
x0 = hslider("x0",0.5,0,1,0.01) : si.smoo;
y0 = hslider("y0",0.5,0,1,0.01) : si.smoo;
y1 = hslider("y1",0,0,1,0.01) : si.smoo;
q = hslider("q[acc: 0 0 -10 0 10]",30,10,50,0.01) : si.smoo;
//////// mapping ////////
impFreq = 2 + x0*20;
resFreq = y0*3000+300;
//////// putting it together ////////
process = os.lf_imptrain(impFreq) : fi.resonlp(resFreq,q,1) :
ef.cubicnl(y1,0)*0.95 <: _,_;
```

Here, `x0` corresponds to the X position of the first finger to touch the screen, `y0` its Y position and `y1` the Y position of the second finger to touch the screen. The q parameter of the resonant lowpass filter is controlled by the X axis of the built-in accelerometer with a linear mapping.[11]

An exhaustive list of the SmartKeyboard configuration keywords can be found in its corresponding documentation[12] and tutorials demonstrating how to implement various types of behaviors can be found on the Faust tutorial page.[13]

6 Digital Lutherie and Smartphones

In parallel of the development of the various tools presented in the previous sections, an important work has been carried out at GRAME-CNCM and at CCRMA[14] (Stanford University) around the concept of augmenting mobile devices to implement advanced musical instruments. The core idea of this project

[11] `https://faust.grame.fr/doc/manual#sensors-control-metadatas`
[12] `https://ccrma.stanford.edu/~rmichon/smartKeyboard/`
[13] `https://ccrma.stanford.edu/~rmichon/faustTutorials/#making-faust-based-smartphone-musical-instruments`
[14] *Center for Computer Research in Music and Acoustics*

was to use mobile devices as the platform for computing and sound synthesis/processing of physical Digital Musical Instruments (DMIs) built around this type of device. Two kinds of "smartphone augmentations" were developed in this context:

- **passive augmentations** [10] based on digitally fabricated elements leveraging existing sensors on the device, allowing us to hold it in a specific way, or modifying the acoustical properties of its built-in speaker and microphone, etc.,
- **active augmentations** [11] implying the use of additional sensors connected to the mobile device through the use of a microcontroller, etc.

Figure 4 presents an overview of the type of passive augmentations that have been explored as part of this project.

Fig. 4. A Few Examples of Passive Smartphone Augmentations.

The BladeAxe [12] is a good example of an active mobile device augmentation. It provides a plucking system based on piezo to capture sound excitations created by the performer on plastic tines to drive waveguide physical models running on an iPad in an app implemented with `faust2smartkeyb` (see §5). This allows for a very natural and intuitive control of the plucking since the sound of each excitation is different.

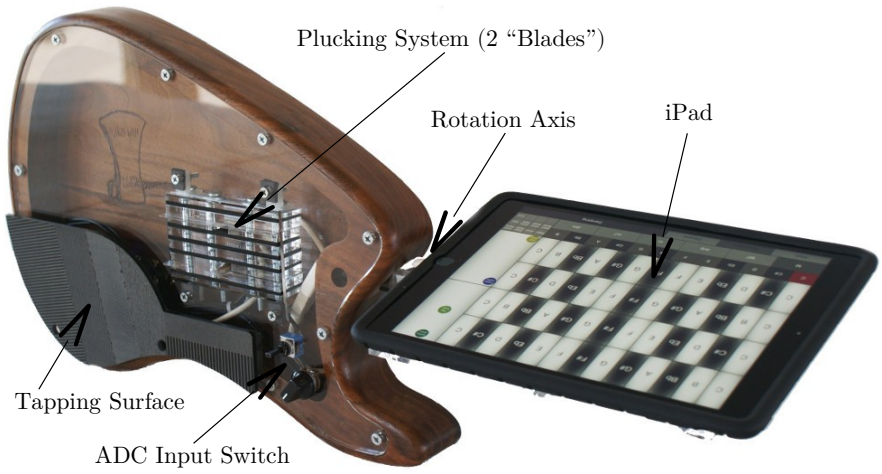


Fig. 5. The BladeAxe.

7 Performances and Pedagogy

7.1 From SmartFaust to Geek-Bagatelles

`faust2ios` and `faust2android` (see §2,3) served as the platform for the development by GRAME-CNCM of a series of large scale musical projects involving the use of smartphones as early as 2013. The first of them – *SmartFaust* – was a participatory concert for smartphones which was commissioned to composer Xavier Garcia for the 2014 Biennale Musique en Scène and funded by the INEDIT ANR[15] project.

Garcia worked closely with a developer/computer music assistant (Christophe Lebreton) to the development of a series of iOS and Android applications using `faust2ios` and `faust2android`. The instruments/applications and their corresponding musical pieces were co-written simultaneously. Another remarkable feature of these instruments is the lack of graphical interface: only motion sensors were used. The performer never needs to look at the phone to play it: everything is done between the hand and the ear!

The fruit of this work was performed for the first time at the Subsistances in Lyon (France) in March 2014. The concert was organized in two sections: the performance of three pieces for “chorus” of Smartphones and soloists, and then a fourth piece involving the audience.

After this first performance, *SmartFaust* met a large success and started an Asian tour with participatory concerts and workshops that were organized in June 2015 in Wuhan, Hong-Kong and Chengdu. In this context, new pieces for

[15] *Agence Nationale de Recherche*: French National Research Agency

Fig. 6. Left: *SmartFaust* Performance at the Subsistances in Lyon (France) on March 16, 2014. Right: *SmartMômes* Performance at the Saint-Étienne (France) City Hall in March 2016.

the *SmartFaust* apps corpus were written, in particular by composer Qin Yi in Shanghai.

The original *SmartFaust* project also gave birth to other performances such as:

- *SmartFaust on Air* at the 2015 Design Biennale in Saint-Etienne (France),
- participatory concerts in the TGV[16] in partnership with the SNCF,[17]
- sound installations with the *Smartland Divertimento* piece presented at the Museum of the Confluences at the 2016 Biennale Musique en Scène (Lyon, France).

The latter, proposed by Christophe Lebreton and composer Stéphane Borrel, is like a bush of smartphones that communicate with each other and sparkle independently, a bit like fireflies.

The most recent project of this series was created as part of the ONE project (Orchestra Network for Europe) with the Picardy Orchestra (France). It was finalized in September 2014, approved by the European Commission in April 2015, and finally resulted in a commission to composer Bernard Cavanna for a piece for orchestra and smartphones: *Geek-bagatelles, introspections sur quelques fragments de la IXe symphonie de Beethoven*. The performance was premiered on November 20, 2016 by the Picardy Orchestra at the Paris Philharmonie. It combined a chorus of 20 smartphones and an orchestra of 38 musicians. The audience participated as well thanks to the *Geek-Bagatelles* app on their smartphone.

The performance was a success and a tour was initiated in the countries part of the ONE network, each time with a new orchestra and a new amateur smartphones chorus formed for the occasion.

[16] High speed train system in France
[17] French National Railway Company

Fig. 7. *Geek-Bagatelles* Performance at the Paris Philharmonie on November 20, 2016.

7.2 SmartMômes

After the initial performance of *SmartFaust*, Mômeludies which is a nonprofit promoting the creation and the diffusion of new musics towards kids commissioned composer Xavier Garcia a new piece for smartphones: *SmartMômes* (see Figure 6). They asked him to teach a series of workshops on this topic in multiple middle schools as well. As a publisher, Mômeludies also published the score of *SmartMômes*.

Because of the interest around the pedagogical aspect of this approach, GRAME-CNCM organized a series of *SmartFaust* workshops during which the FaustPlayground[18] was used to create musical smartphone apps using Faust at a very high level with a Graphical User Interface.

8 Current and Future Directions

The various tools and technologies presented in the previous sections of this paper reached a certain level of maturity and are now broadly used at GRAME-CNCM and elsewhere. They significantly contributed to the success of most of the recent musical productions of our center thanks to their universal aspect and to their tangibility. Performing with independent standalone and tangible DMIs is quite appealing in a world where everything tends to become completely virtual. Hence, while we keep adding new features to our toolkit for mobile development, we also started exploring new paths to work with embedded systems for low latency/high quality audio. Indeed, microcontrollers are now powerful enough to run complex sound synthesis and processing algorithms in real-time. Similarly, embedded computers such as the Raspberry Pi (RPI) when used without operating system ("bare-metal"), FPGAs[19], GPUs[20] and other low-level

[18] `https://faust.grame.fr/faustplayground`
[19] *Field Programmable Gate Arrays*
[20] *Graphical Processor Units*

DSPs offer new possibilities to create embedded/embodied instruments at a low cost and with un-paralleled performances. While we currently investigate the use of FAUST on FPGAs and bare-metal RPI, FAUST targets have already been implemented for microncontrollers [13] and DSPs such as the SHARC Audio Module.[21]

These new developments recently allowed us to create a new programmable musical instruments: the *Gramophone* (see Figure 8) that we plan to use for pedagogical purpose and for future musical productions at GRAME-CNCM. Based on Teensy 3.6 board[22] for sensor acquisition and sound synthesis, it can be powered by its internal battery for about ten hours, it is equipped with a powerful speaker and amplifier, and it hosts a wide range of sensors (i.e., accelerometer, gyroscope, compas, force sensing resistors, knobs, buttons, photo-resistor, etc.) that can be assigned to FAUST parameters directly from the FAUST code using metadata. It is better than a smartphone in many ways as it offers more affordances and it is more flexible and much louder. While it is still being developed, we plan to release the first version in Fall 2019.

Fig. 8. The Gramophone.

9 Conclusion

After fifteen years, mobile music has reshaped the computer music landscape partly by reintroducing the concept of standaloneness/independence in DMIs and by making this type of instrument more approachable by the general public. FAUST played a role in this revolution by providing high level tools to develop musical apps for live performance. GRAME-CNCM took advantage of these technologies to place mobile music at the heart of various large scale musical

[21] `https://wiki.analog.com/resources/tools-software/sharc-audio-module/faust`

[22] `https://www.pjrc.com/store/teensy36.html`

productions/projects. By offering the possibility to easily create orchestras of DMIs, mobile music opened the way to new paths for creation that we intend to keep exploring by developing new programmable instruments taking advantage of recent developments in embedded real-time signal processing such as the Gramophone.

References

1. Gaye, L., Holmquist, L.E., Behrendt, F., Tanaka, A.: Mobile Music Technology: Report on an Emerging Community. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME-06), Paris (2006)
2. Tanaka, A.: Mobile Music Making. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME04), National University of Singapore (2004)
3. Schiemer, G., Havryliv, M.: Pocket Gamelan: Tuneable Trajectories for Flying Sources in Mandala 3 and Mandala 4. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME06), Paris (2006)
4. Wang, G.: Ocarina: Designing the iPhone's Magic Flute. Computer Music Journal, 38(2), 8–21 (2014)
5. Orlarey, Y., Letz, S., Fober, D.: New Computational Paradigms for Computer Music, chapter "Faust: an Efficient Functional Approach to DSP Programming." Delatour, Paris (2009)
6. Michon, R.: faust2android: a Faust Architecture for Android. In: Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland (2013)
7. Michon, R., Smith, J.O., Orlarey Y.: MobileFaust: a Set of Tools to Make Musical Mobile Applications with the Faust Programming Language. In: Proceedings of the International Conference on New Interfaces for Musical Expression, Baton Rouge (2015)
8. Michon, R., Smith, J.O., Letz, S., Chafe C., Orlarey, Y.: faust2api: a Comprehensive API Generator for Android and iOS. In: Proceedings of the Linux Audio Conference (LAC-17), Saint-Étienne, France (2017)
9. Michon, R., Smith, J.O., Chafe, C., Wang, G., Wright, M.: faust2smartkeyb: a Tool to Make Mobile Instruments Focusing on Skills Transfer in the Faust Programming Language. Proceedings of the International Faust Conference (IFC-18), Mainz, Germany (2018)
10. Michon, R., Smith, J.O., Wright, M., Chafe, C., Granzow, J., Wang, G.: Passively Augmenting Mobile Devices Towards Hybrid Musical Instrument Design. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME-17), Copenhagen (2017)
11. Michon, R., Smith, J.O., Wright, M., Chafe, C., Granzow, J., Wang, G.: Mobile Music, Sensors, Physical Modeling, and Digital Fabrication: Articulating the Augmented Mobile Instrument. Applied Sciences, 7(12), 1311 (2017)
12. Michon, R., Smith, J.O., Wright, M., Chafe, C.: Augmenting the iPad: the BladeAxe. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME-16), Brisbane, Australia (2016)
13. Michon, R., Orlarey, Y. Letz, Y., Fober D.: Real Time Audio Digital Signal Processing With Faust and the Teensy. In: Proceedings of the Sound and Music Computing Conference (SMC-19), Malaga, Spain (2019) – Paper not published yet but accepted to the conference

COMPOSITES 1: An Exploration into Real-Time Animated Notation in the Web Browser

Daniel McKemie

Independent Author
daniel.mckemie@gmail.com

Abstract. *COMPOSITES 1 for Modular Synthesizer Soloist and Four Accompanists* is a real-time, graphically notated work for modular synthesizer soloist and four accompaniment parts that utilizes the power of Node.js, WebSockets, Web Audio, and CSS to realize an OS-agnostic and web-deliverable electroacoustic composition that can be accessed on any device with a web browser. This paper details the technology stack used to write and perform the work, including examples of how it is used compositionally and in performance. Recent developments in web browser technology, including the Web Audio API and Document Object Model (DOM) manipulation techniques in vanilla JavaScript, have improved the possibilities for the synchronization of audio and visuals using only the browser itself. This paper also seeks to introduce the reader to the aforementioned technologies, and what benefits might exist in the realization of creative works using this stack, specifically regarding the construction of real-time compositions with interactive graphic notations.

Keywords: JavaScript; Node.js; WebSockets; CSS; Animated Notation; Web Browser; Web Audio API; Mobile Device Music

1 Introduction

During my studies with John Bischoff and Chris Brown [1], I developed my aesthetic of network music that is largely influenced by West Coast Experimentalism. However, it is only recently that have I made serious developments in my own work through the use of web-based technologies, and in particular, the construction and realization of network-based pieces and real-time compositions for live performance. The work of Georg Hajdu, most notably for this context his Quintet.net [2, 3], has served as a great example of contemporary approaches to network music; bringing the traditions of tethered, machine-based network music [1] into the age of mobile devices and wireless capabilities [4, 5].

2 Full Stack Web Technology

JavaScript is the language of the web, and over the last decade it has expanded greatly into server-side tools and technology. The Document Object Model (DOM) is used to create dynamic changes to, and interactions between, HTML elements on the webpage, and can be linked to a number of processes afforded to work in the browser. For the realization of *COMPSOITES 1*, there is a great degree of nested communication in Node.js using WebSockets, with the Web Audio API treating audio data on the front-end, all linked through JavaScript as the primary mode of construction. This architecture is very similar to that implemented in the *Soundworks* framework built by Sébastian Robaszkiewicz and Norbert Schnell at IRCAM [6]. In this section, I will briefly give an overview of these primary components used to write *COMPOSITES 1 for Modular Synthesizer Soloist and Four Accompanists*.

2.1 Web Audio API

The Web Audio API (Application Programming Interface) is a high-level JavaScript API that enables audio synthesis and digital signal processing (DSP) in the browser. The dynamic nature of JavaScript allows the API to be used in conjunction with an array of libraries and tools available in web development, including those on the server side with Node.js. The structure of the API, specifically the modular nature of audio node routing, is similar to other audio software environments in that it provides the user with a large array of options for the synthesis and processing of audio.

2.2 Node.js

Node.js is an open source server environment that allows JavaScript to run on the server and return content to the client. This is used to power data flow from back end databases all the way to the browser page using a single language. Global modules, variables, and functionalities can be spread across multiple pages in the server and can be used with a multitude of frameworks to maximize the efficient construction of an API. In this case, the popular framework Express was used as it is lightweight and wide support.

2.3 WebSockets/Socket.IO

WebSocket technology allows for real-time data transfer to and from the server, without the need to refresh the webpage. This enables the manipulation of back end data through client activity, the broadcasting of unique front-end HTML/CSS stylings to multiple devices or individuals, and the projection of real-time manipulations of said stylings to multiple URLs.

Socket.IO is a client and server-side library that uses WebSockets, with some added features that can open thousands of connections vs other methods [7]. Socket.IO is optimized to implement real-time binary streaming, and for this reason was chosen as the best option for the needs of this project.

3 COMPOSITES 1

In the musical work *COMPOSITES 1 for Modular Synthesizer Soloist and Four Accompanists*[1], there is one host point (the soloist) who goes to the homepage/server site, hooks a modular synthesizer in to be the performance instrument, and assists in score generation. Each of the four accompanying parts can be assigned to any pitched instruments and can be augmented with electronics if desired. The soloist is instructed to send out a simple waveform from the synthesizer, with sine, triangle, and sawtooth working best in that order, into the computer (via an audio interface) which will then send the signal to the web browser to be pitch-tracked. The resultant data of the pitch will be used to generate the notational material for the accompanists. While this waveform is uninterrupted in its signal path directly to the server site, the soloist is to construct a performance patch around that patch point before it reaches the server. The performer may modulate that waveform's frequency but is advised not

[1] Demonstration video: https://www.danielmckemie.com/composites1

modulate too heavily: a "cleaner" tone, such as a sine wave, will lead to a more legible performance score for the accompanists.

The frequency is tracked, and certain frequency thresholds lead to the generation of visuals and/or "decisions" to change the existing visuals. The visuals include color changes, projected pitches on the staff, shapes, shade contours, and so on. These visuals are broadcast to individual URLs, and each of the four accompanists can then access on their own device. For the sake of brevity, I will only discuss examples of one accompanying part, and the technology used to create the server and client pages.

3.1 File Structure

The file directory is set up with client-side scripts in the public folder that are sent to the browser screen which is nested inside the working directory of the Node.js and Socket scripts. Beginning with the client-side calls, there is an HTML and JavaScript file for the host and each accompanist respectively. The host file holds the connections to the server, sockets, and analysis code for the incoming audio stream.

3.2 Client-Side Input and Analysis

An audio input stream is created using the Web Audio DAW (Wad) library [8]. Based on the same concept as a guitar tuner, the signal's frequency is tracked, and because the signal is a simple waveform from an electronic source, the stability of tracking is far better than that of an acoustic instrument (Fig. 1). A major benefit of the Web Audio API is that the hardware synchronization is all done globally. The audio configurations in the computer's system preferences are automatically picked up by the browser, which frees the need for any external drivers or OS dependencies.

The code in Fig. 1 calls for the input and runs the signal through the logPitch() function which is a product of the Wad library. This function calls to analyze the frequency of the incoming signal, and also return its corresponding note name as it coincides on the staff.

```
// host.js

let input = new Wad({ source: 'mic' });
let tuner = new Wad.Poly();

// Sets the library to begin tracking input signal info
tuner.setVolume(0);
tuner.add(input);
input.play();
tuner.updatePitch();

// Logs the signal to frequency number and note name
let inputFreq = null;
let inputNote = null;
let logPitch = function() {
  requestAnimationFrame(logPitch);
  inputFreq = tuner.pitch;
  inputNote = tuner.noteName; };
```

Fig. 1. Code to declare input and track frequency. The frequency and note name captured by the logPitch() function is assigned to the global inputFreq and inputNote, to allow for transfer and broadcast to the clients.

4 Client to Server Connection

4.1 Client-Side

After the input, analyses, and assignments on the client-side have been rendered, the WebSockets must connect this information to the server in order for it to be broadcast to the client's webpages. The 'broadcast' button calls for the host data to be emitted via the socket, and back to the server, and this occurs automatically every 20 milliseconds in order to simulate the transmission of real-time data flow. A number of sockets can be implemented to send data, and each are identified with unique IDs created by the user. In the following case 'frequency' is the unique ID associated with the analysis of the incoming audio stream (Fig. 2).

```
// host.js
// A trigger that automates a button click every 20ms
const buttonBroadcast =
document.getElementById('broadcaster');
setInterval(function() {
  buttonBroadcast.click()}, 20);

// Clicking the button sends the pitch data to the server
buttonBroadcast.addEventListener('click', function(e) {
  e.preventDefault();
  socket.emit('frequency', {
buttonBroadcast.addEventListener('click', function(e) {
  e.preventDefault();
  socket.emit('frequency', {
    pitch: inputFreq,
    note: inputNote })
});
```

Fig. 2. The frequency assignment sends the pitch and note keys with the values of the global variables, inputFreq and inputNote, which were declared prior. This sends/emits this data to the server environment via Socket.IO

4.2 Server Side

The server connection to the localhost and the incoming sockets are all housed within the primary application file in Node.js. Following the thread from the client, the socket connections must be written in such a way that the host data can pass to anywhere on the server via Socket.IO. In this case, the data is broadcast to all clients on the server; and while there are options to broadcast to select clients instead of broadcasting all data in one stream (frequency) while avoiding others, it is beyond the scope of this paper, and not necessary for the success of the piece (Fig. 3).

```
// app.js
io.on('connection', function(socket) {
```

```
  socket.on('frequency', function(data) {
    io.sockets.emit('frequency', data) });
```

Fig. 3. This code enables global data sharing and accessibility among all the clients but must be called later by specific clients for their own use.

4.3 Return to Client

The data received from the host and brought to the server, can now be pulled back through to any of the four remaining clients (accompanists) through their respective script.js files. In order to see unique transmissions and treatments of this data, a route must be set up in our app.js file so that a connector can access for their own broadcast (Fig. 4).

The JavaScript calls the data emitted by Socket.IO and can be treated in any number of ways. In this example, the frequency number is run through a switch statement, and as certain conditionals are met, the background color of our <body> is changed accordingly (Fig. 5). This data can be sent to clients in many fashions, from private messages to select clients, and so on, but for the sake of brevity we will emit data to all clients equally.

```
// app.js
// Piping the pitch data back up to the client-side page
app.get('/player1', function(req, res) {
  res.sendFile(__dirname + '/public/player1.html');
});

// script1.js
// As the data is read, a connection is made between the
// background color of the web page and the pitch
// material. A switch statement changes the colors
// according to frequency range

let backColor = null;
socket.on('frequency', function(data) {
  switch (true) {
    case (data.pitch < '200'):
      backColor = 'red';
      break;
    case (data.pitch < '300'):
      backColor = 'orange';
      break;
   };
  document.body.style.backgroundColor = backColor });
```

Fig. 4. The app.js file sends our player1.html file to the browser, which is linked to our script1.js file, which can treat the data sent out to the server through WebSockets.

5 Use as a Compositional Tool

The browser houses a healthy number of ways to compose, manipulate, and animate visual elements, and CSS and DOM manipulation techniques alone include a large number of options to create scores that convey musical information to performers. As outlined in previous sections, these elements can be synchronized through the server and delivered to individually unique locations. The following will outline a number of examples as to how *COMPOSITES 1* uses these techniques to create a real-time score for the accompanying parts.

The elements of the score include the background color, shapes, words and staff notation. Changes to these elements are based on the host input's frequency and/or note name values. The goal of the work was to not inundate the performers with instructions and elements, but rather let them choose pathways in which to realize musical material. This was to lessen the problems that can arise when constructing notation in real-time. [9]

5.1 Colors

The use of color as a notational element has seen more traction in recent years, but no consistent practice has been established. Lindsay Vickery's research into the topic suggests the need for further exploration into the field, especially in regard to multimedia works [10]. As opposed to leaving the interpretation of colors open or to be designated for assignment, the decision was made to give the performer overall qualities to enhance the stylistic qualities of their playing. A classic and notable example of color usage is that of John Cage's Aria (1958), in which he uses color to denote different singing styles for each line [11]. I wanted to take a similar approach in *COMPOSITES 1*, to denote a type of executable action in response to color as opposed to a specific executable action. The colors of the spectrum are at play, and call for the following:

- Red: Poignant
- Orange: Fleeting
- Yellow: Bright
- Green: Stable
- Blue: Metastatic
- Indigo: Dark
- Violet: Razor-like

Fig. 5. Example of the frequency to color relationship as seen through the host's console (right) and Player 1's yellow background (left).

5.2 Shapes

Shapes have been a centerpiece in the evolution of western music notation and maintain an important role in certain brands of music education and vocal music [12]. While I did not seek to morph and modulate the current staff notation system, I did want to use basic shapes to convey a musical response, but not have any shapes that were simply a one-dimensional linear contour. This decision stems from not wanting to encourage a correlation of pitch with vertical placement [10], though if players wish to interpret the two-dimensional shapes this way, that is quite acceptable. With the circle, triangle, and square, the instruction to the performer is to assign a very focused action that best reflects the given shape and execute it regardless of surrounding context. Simple shapes were chosen as they can be easily recognizable compared to the others, and should result in a more focused performance [13]. This achieves two desired goals, the first being a consistent execution of sonic events that will occur throughout the piece; and the second, to allow the performer to have more unrestricted decision making in the performance process. The shapes are called by a change in the <*img*> tag source path and are set to appear when specific parameters are met.

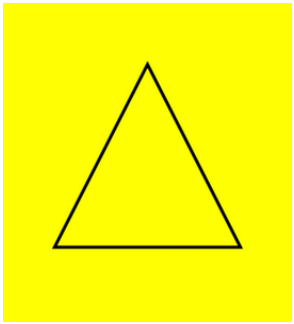


Fig. 6. The triangle shape used in COMPOSITES 1 (one of three shapes to appear), with the yellow background.

5.3 Words

The use of written words to enhance musical expression and performance has been around for centuries. In the case of *COMPOSITES 1*, instead of indicating musical results that could be achieved by using traditional markings (ie. *ff* instead of 'very loud' or a crescendo instead of 'get louder'), I sought to explore the use of words to allow for variations upon the already present situation. To expand, when 'pointed' is displayed, this can result in a number of different outcomes based on both decisions made by the performer, and the context in which it appears (Fig. 7).

The words are generated independently from the host's incoming audio signal by a random number generator and a switch statement assigned to the DOM. Inserting one independently-timed element that is separate from the host's actions gives the accompanying part its own pace. This decision grew out of a desire to eliminate any sort of perceptible rigidity in the performing group as a whole, without having completely asynchronous events.

Additionally, the use of simple, non-musical phrases enhances the space that the performer can work in, and at the same time not be overbearing. The words and shapes are the only elements of the score that appear and are then removed as opposed to remaining stationary and changing over time.

5.4 Pitch/Staff Notation

The images of the staff are the most straightforward of all the elements in the score. As the note changes, the performer uses it as a pitch reference to construct their material, and they also have the option to ignore any number of other elements and continue holding that pitch. The note indicated on the staff is directly correlated to the note name of the incoming waveform.

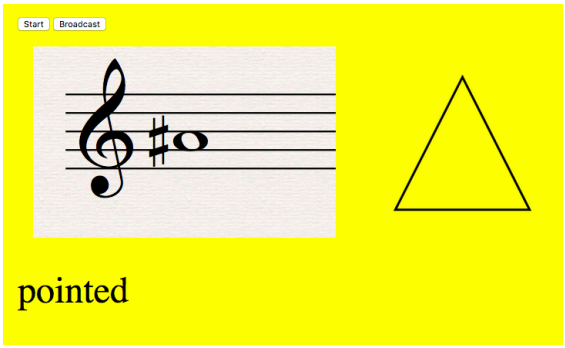


Fig. 7. An example of all elements displayed on one page as seen by an accompanying performer.

5.5 Goals and Challenges

As stated earlier, the resulting notation structures are designed to produce varying interpretations by the accompanying performers. The instructions provided encourage the performer to choose pathways and focus on or ignore elements as they choose. Granting the performers this flexibility helps to relieve any desire, whether purposeful or accidental, to simply follow the soloist. The use of colors, shapes, words, and traditional staff notation provides varying degrees of openness to closedness, and players can find their own level of comfort in the notational structures provided.

The chaotic nature of analyzing audio signals for pitch detection becomes potentially problematic for the stability of the display of accompaniment scores. For example, an input signal that accidentally becomes frequency-modulated would in turn send the accompanists' notation screens into a tail spin, with elements rapidly changing well

beyond any realistic readability. Instead of leaving this open to cause an otherwise well executed performance to be ruined, this has been dealt with in two ways: firstly, the soloist is instructed to extensively practice with their synthesizer and this setup and to take note of reactions based on their actions; and second, if the notation system were to fall into chaos, accompanying performers are informed to do the same. Instead of treating this as a hinderance, the inherent fragility in the technology becomes an interesting musical element when executed properly.

6 Future Work

The ever-expanding collection of libraries and development tools can be seamlessly integrated into and inspire works like *COMPOSITES 1*. More involved data visualizations can be built using libraries like D3 [14], and the sharing of data over common networks to control and manipulate individual Web Audio streams could lead to very interesting results, as it did in similar settings that came before it [1].

The back-end server is what supports the delivery of the parts to individual screens for each player and allows for a greater depth of concentration and execution of each player's part for a more effective performance. Taking inspiration from Kelly Michael Fox's *Accretion* [15], which uses individual monitors as opposed to a projected screen in which all players read from, not only maximizes the efficiency of the space in which to render elements, but also injects a sense of mystery for the audience: what are they all looking at?

The fact that the projected parts can be accessed via a multitude of different devices, so long as they support modern web browsers, is key in broadening accessibility and ease of performance of works such as *COMPOSITES 1* and those that choose to employ the same architecture. As stated earlier, the Web Audio API uses the global settings of the computer, which allows for a wider range of device types to be used, and the option to deploy the entire piece as a cloud-based app [16], removes the need to deliver code in any form, requiring nothing more than a simple URL.

7 References

1. Bischoff, J., Brown C.: Indigenous to the Net: Early Network Music Bands in the San Francisco Bay Area, http://crossfade.walkerart.org/brownbischoff/IndigenoustotheNetPrint.html (2002)
2. Hajdu, G.: Quintet.net: An Environment for Composing and Performing Music on the Internet. In: Leonardo, vol. 38, num. 1, pp. 23-30 (2005)
3. Hajdu, G.: Real-time Composition and Notation in Network Music Environments. In: Proceedings of the International Computer Music Conference. Belfast, N. Ireland (2008)
4. Hajdu, G.: Composing for Networks. In: Proceedings of the Symposium for Laptop Ensembles and Orchestras, pp. 98-102. Baton Rouge, Louisiana, USA (2012)
5. Carey, B.: SpectraScore VR: Networkable virtual reality software tools for real-time composition and performance. In: International Conference on New Interfaces for Musical Expression, pp. 3-4. Brisbane, Australia (2016)

6. Robaszkiewicz S., Schnell N.: Soundworks – A playground for artists and developers to create collaborative mobile web performances. In: 1st Web Audio Conference. Paris, France (2015)

7. Carey, B., Hajdu, G.: Netscore: An Image Server/Client Package for Transmitting Notated Music to Browser and Virtual Reality Interfaces. In: The International Conference on Technologies for Music Notation and Representation, pp. 151-156. Cambridge, UK (2016)

8. Serota, R.: Web Audio Library, https://github.com/rserota/wad

9. Freeman, J.: Extreme Sight-Reading, Mediated Expression, and Audience Participation: Real-Time Music Notation in Live Performance. Computer Music Journal, vol. 32 no. 3, pp. 25-41 (2008)

10. Vickery, L.: Some Approaches to Representing Sound with Colour and Shape. In: The International Conference on Technologies for Music Notation and Representation, pp. 165-173. Montreal, Canada (2018)

11. Poast, M.: Visual Color Notation for Musical Expression. Leonardo, vol. 33, no. 3, pp. 215-221 (2000)

12. Johnson, D.C.: Tradition with Kodály Applications. Kodály Envoy, vol. 35, no. 1, pp. 11-15 (2008)

13. Smith, R.R.: An Atomic Approach to Animated Music Notation. In: The International Conference on Technologies for Music Notation and Representation, pp. 40-48. Paris, France (2015)

14. Data-Driven Documents, https://d3js.org

15. Fox, M. K.: Flexible, Networked Animated Music Notation for Orchestra with the Raspberry Pi. In: TENOR: The International Conference on Technologies for Music Notation and Representation, 104-109. Paris, France (2015)

16. Deploying Node.js Apps on Heroku, https://devcenter.heroku.com/articles/deploying-nodejs

Realtime collaborative annotation of music scores with Dezrann

Ling Ma, Mathieu Giraud, and Emmanuel Leguy

Univ. Lille, CNRS, Centrale Lille, UMR 9189 – CRIStAL
Centre de Recherche en Informatique Signal et Automatique de Lille
F-59000 Lille, France
dezrann@algomus.fr

Abstract. Music annotation is an important step in several activities on music transcribed in common music notation. We propose a protocol to annotate collaboratively such scores in real time. Based on a paradigm with commutative operations, this protocol guarantees consistency beween distributed editions while providing a fluid user experience, even behind possible network lags. It is being implemented into Dezrann, a web platform for sharing music analysis. We report efficiency and scalability tests on the current implementation, including usage by up to 100 simulated clients.

Keywords: Score annotation, music analysis, collaborative editing, distributed algorithms, operational transformation, operation commutation

1 Introduction

Music annotation is an important step in several activities on music transcribed in common music notation – music analysis, teaching, performance preparation or even composing. As we noted in [5], annotation can be modeled as putting *labels* on the score at various positions, possibly with a duration, and sometimes drawing relations between these labels. Researchers in musicology or in computer musicology often engage with scores through reading, annotating, and analyzing, and they discuss other people's analyses. In talks or lectures on music analysis, history or composition, one frequently needs to *get back to the score*, jumping to different sections, sometimes comparing them, in different places, commenting some elements of a score or of several scores. Most of the time, *paper scores* are very efficient for that. In such lectures, it is now common to see students or people with laptops or tablets. But they often are not as efficient as when they work with annotated paper scores laid on the desk. More generally, many people like to talk about music and to describe what they hear, as demonstrated by presence of music on social networks.

Software for music annotation. Can software help people from different backgrounds efficiently annotate or discuss music? Several software enable to render, and sometimes to annotate, scores on the web[1] [7, 13, 14], or to analyze music [3,

[1] jellynote.com, notezilla.io

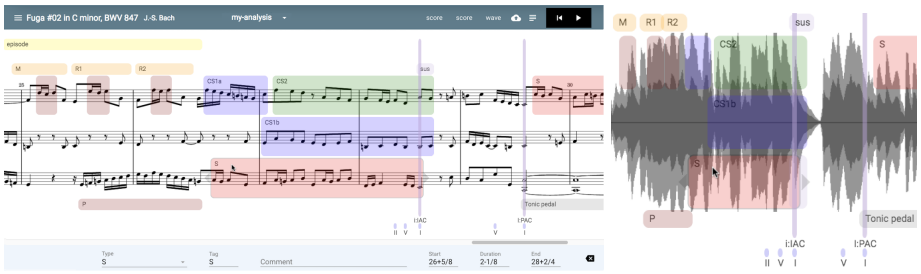


Fig. 1. Annotation with Dezrann (`www.dezrann.net`) on a fugue in C minor by J.-S. Bach, showing both on-stave (subjects, counter-subjects, patterns) and off-stave (harmonic sequence, degrees, cadence, pedal) labels. The music/label synchronization is kept across both a generated score (left) and a waveform from a recording (right).

8, 10]. We introduced the open-source platform Dezrann [6] favoring simplicity over the number of features, focusing on easy usage for people with limited programming background as well as efficiency for people needing to encode music annotations on large corpora. Music is presented either on continuous staves or on waveforms. The user may create, edit and move *labels* on the score – either on a staff or on spaces above or under the score (Figure 1).

Realtime collaborative annotation. Collaborative editing is made possible by fast and ubiquituous network connexions. Initially proposed to edit texts, it is also used in graphics or other domains. Collaborative music applications on the web began to emerge a few years ago, with for example collaborative score edition[2], collaborative patching [12], or collaborative performance [1, 11].

Behind the scene, the challenge of any collaborative editor is to maintain *consistency* between multiple editors separated by the network, and to handle conflicts possibly arising from simultaneous editions. *Lock* mechanisms are reliable, but they limit real-time interaction. More recent models enable actual simultaneous editions by several clients. Operational Transformation (OT) is based on transforming the operations to keep a converging state for each client [4], keeping casual links between operations – one operation possibly being emitted after another one. Decentralized protocols using conflict-free replicated data types (CRDT) guarantee that the user intention is preserved by using only operations that are associative, commutative and idempotent (applying them once or several times yield the same result) [15].

Aims and Contents. Considering the needs of collaborative music annotation, and taking ideas from existing techniques, we propose a distributed protocol to handle simultaneous editions on a set of *labels* on traditional scores. The next Section details our motivations and use cases, Section 3 describes the proposed protocol, Section 4 presents the implementation and the evaluation, and Section 5 concludes on the availibility and the perspectives.

[2] `flat.io`

Fig. 2. Children using Dezrann in a public secondary school in Amiens (France) to annotate sections in music. The school curriculum in music education in France includes a part of "music analysis", without assuming that the children are music readers. Having a way that children annotate themselves music make them active. It improves their learning, and more generally their autonomy.

2 Annotation Model and Use Cases

Annotating music can take different ways, even when one considers only scores in common music notation. We use here the simple notion on *labels* on the score as defined in [2]. Each label has several *fields*. It has a start onset, it may have a duration, or not, and it may concern the whole system, or, more precisely, one or several staves. Common labels to analyze tonal music are patterns, cadences, harmonic markers, or structural elements. This simple modeling does not cover all aspects: Most notably, people analyzing music on paper frequently draw connexions between labels. Here relations between labels are not modeled, but nevertheless labels can have tags or comments. Within this simple modeling, collaborative edition could be worth both in colocated and remote situations:

- *Event-based interactive collaboration.* People, usually in the same place, discuss music while annotating it on their computers, tablets or other mobile devices. They see how the other people interact and annotate the score, for example in the following situations:
  - *specialized music teaching* (general music teaching, or more specialized lectures such as composition, analysis, or history): Students can learn notions such as the sonata form by annotating scores collaboratively;
  - *music education:* Without detailed analysis on the score, music culture and theory can be shared and experimented, for example by identifying sections and other striking events on a waveform (Figure 2);
  - *conference or public concert:* Audience feedback could be organized by letting people annotate the music.
- *Remote collaboration.* People in different places may, over the network, discuss on a score or a waveform. This could be a few collaborating people – artists preparing a concert, or scholars working on scores – or many more people, in a *social network* to annotate music. Like on platforms like *soundcloud*, people may want to share what they hear and comment music.

These use cases are not exclusive: Collaborative edition makes it possible to host *distributed events* on the network. Moreover, one could also *combine "real-time" and "offline" parts*: A collaborative annotation done during a lecture or

another event could be later available for remote viewing or editing. Conversely, artists or students could prepare their own annotation of a score, and thereafter share and edit collaboratively this annotation with others, either to a group of identified users or towards everyone.

Note also that the collaborative edition could be *unrestricted*, where every user may annotate various places in the music. However, one could also have mechanisms to *highlight* the edition or the interaction of one particular user, as a teacher or a presenter. This could include the ability to *follow the session* of such an user, in particular by browsing the music at the same position of her.

3 Operations Modeling and Network Protocol

The following paragraphs describe how we model the collaborative edition between several *user clients* and a *server*. OT and CRDT bring some interesting ideas as the preservation of *casual links* and the *user intention*, the fact that the operation should be applied in any order or the idempotency operations. However, transforming operation in OT is a complex task, and OT modeling is sometimes error-prone [9]. Moreover, in our scenarios, it makes no sense to merge some conflicting operations such as when two users try to move the same label both to the left and to the right. One of these users should know as soon as possible that he is conflicting and then rejoin the shared state. We thus propose a simpler, centralized model based on a centralized linear history but with "Operation Commutations". Compared with OT, we use only trivial transformations on commutative operations, allowing them to be delayed. The consequence is that not all operations are accepted by the server, leading to limited conflicts that may be easily handled by the concerned clients.

State and operations. On a given score, the *state* of the annotation is the result of the application of an *operation history* onto the State 0 – without any labels. We consider three operations: *label-create(id, dict)* creates a label, *label-remove(id)* removes a label, and *label-set(id, dict)* sets some fields of an existing label.

Label *id*s are computed by hashing and are supposed to be unique. The *dict* arguments are dictionnaries containing one or several *(field, value)* pairs. For example, a label may be created by *label-create(5x8, {start: 10, duration: 4})* and another label can be updated by *label-set(d9t, {start: 36})* or *label-set(d9t, {type: 'Cadence'})*.

Operation commutativity. We consider that two operations performed by different users *on different labels* do commute, because labels are independent objects, without casual links. We also consider that, on a same label, two operations *on different fields* do commute. This allows to be very flexible while keeping the server simple. The *label-set(d9t, {start: 36})* and *label-set(d9t, {type: 'Cadence'})* operations do thus commute, enabling to simultanously edit the same label *d9t*, as on Figure 3. In fact, the server is not aware of the actual semantics of the operation: To check whether two operations on a same label commute, it just checks whether they concern different fields or not.

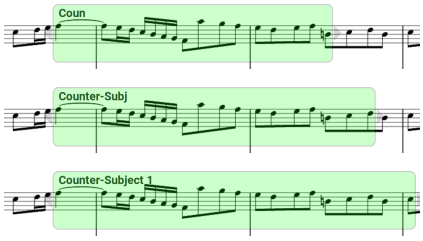


Fig. 3. Andreas is taking a few seconds to enter the comment of a label, here `Counter Subject 1`. One operation is sent by his client at each character stroke. While this, Beatrix is slighlty updating the end of that pattern, resulting again in several operations while she uses her mouse. Her operations do commute with Andreas' operations. Both of them can thus simultanously edit the label.

Client-server handshaking. The server remembers a *state number* s_0 and an *history*[ ] array of size s_0 with all *accepted operations*. User clients may request at any time this full history, in particular when they join the collaborative channel. Each client i, with $i \geq 1$, keeps his own *state* s_i reflecting the last acknowledgement (OK) he received from the server, and a list of his *pending operations* that were sent to the server since this s_i. A client wanting to do an operation op_k sends (s_i, op_k) to the server (Figure 4). When the server receives such a pair:

- If $s_i = s_0$, the client i was aware of the last accepted operation. The server accepts the operation op_k, meaning that he increments $s_0 \leftarrow s_0 + 1$, then stores $history[s_0] = op_k$, and broadcasts to each client $\text{OK}(s_0, op_k)$;
- If $s_i < s_0$, the client i was not aware of the accepted operations since $history[s_i]$. The server checks whether op_k commutes with all $history[s_i + 1] \ldots history[s_0]$ operations that were not from the same client:
  - If the operation commutes (or if there were no operations from other clients), then op_k is accepted *at a new position*, that is $s_0 \leftarrow s_0 + 1$, $history[s_0] = op_k$, and broadcast $\text{OK}(s_0, op_k)$;
  - In the other cases, the server sends $\text{DENY}(op_k)$ only to the client i.

When a client receives $\text{OK}(s_0, op_k)$ from the server, he updates his s_i. When op_k was his own operation, he removes it from his list of pending operations. Otherwise, when op_k is commuting with his pending operations, he applies it ($\star$ on Figure 4). When this is not the case ($\bullet$ on Figure 4), or when he receives a $\text{DENY}(op_k)$, the client detects a conflict and rollbacks some operations, possibly asking some of the *history*[ ] to the server. Thus when two or more users are in conflict, some of them will be blocked, but it can be acceptable.

The protocol allows to detect lost messages: Whenever a client receives $\text{OK}(s_0, \ldots)$ with $s_0 > s_i + 1$, he can ask again some of the *history*[ ] to the server. As the history is linear and as the only transformation is to switch commutative operations, the protocol tries to preserve some form of casual links and user intention. Note that all the update operations are idempotent, and could thus applied several times in case of re-emission following network failures.

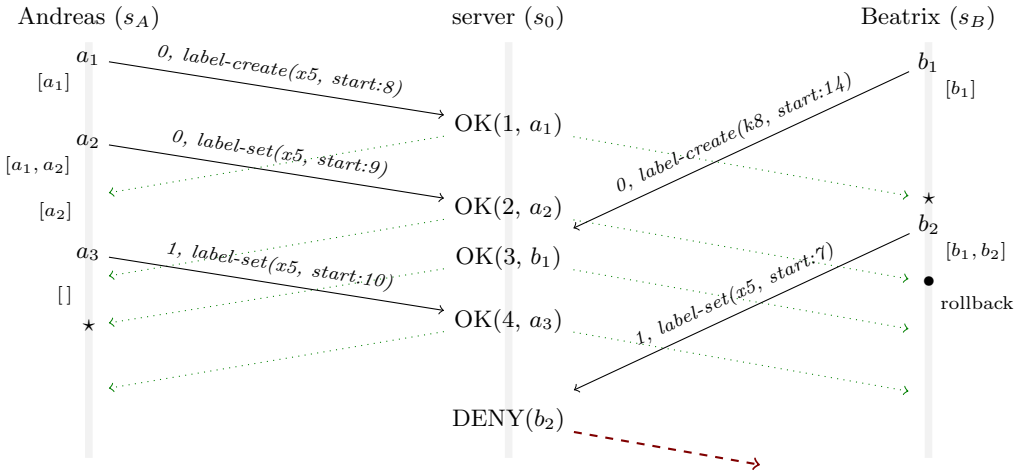


Fig. 4. Example of protocol based on Operation Commutations for collaborative music annotation. Andreas emits operations a_1 and a_2. The server accepts these operations, stores them in *history*[], and broadcasts the acknowledgements. Each time Andreas receives the acknowledgements, he update his s_A with s_0 and, when the concerned operation was from him, he deletes it from his list of pending operations. Beatrix emits the operation b_1 = *label-create(k8, start:14)*, without being aware of operations a_1 and a_2. However, b_1 commutes with a_1 and a_2, and thus is accepted, and a_3 is also accepted because it commutes with b_1. Contrarily, the further operation b_2 = *label-set(x5, start:7)*, where Beatrix want to move the label *x5* in a conflicting way, does not commute with a_2 nor a_3. The server sends thus a DENY back to Beatrix. However, Beatrix could have detected the conflict before, when she received OK$(2, a_2)$ at (•).

4 Implementation and Evaluation

4.1 Implementation

The Dezrann codebase includes several user interface parts, using the *Web components* model (`Polymer.js`), that will be not discussed here [6]. The server part, written in `node.js`, handle user authentication through JWT (json web tokens) and storage of corpora and annotation files, together with their metadata. The collaborative annotation protocol was implemented above `socket.IO`. The payload of each operation message is (s_i, op_k), as explained above, together with other data such as hashes to uniquely identify operations.

The conflict handling is now not optimized on the client side: The clients do not detect by themselves the conflicts, but rather wait for DENY messages from the server. Moreover they request the full state after each conflict, and do not check for lost messages. The protocol still works with such unoptimized clients, except when messages are lost. To benchmark the server part, we both tested the actual client, but also simulated fake clients running regular requests.

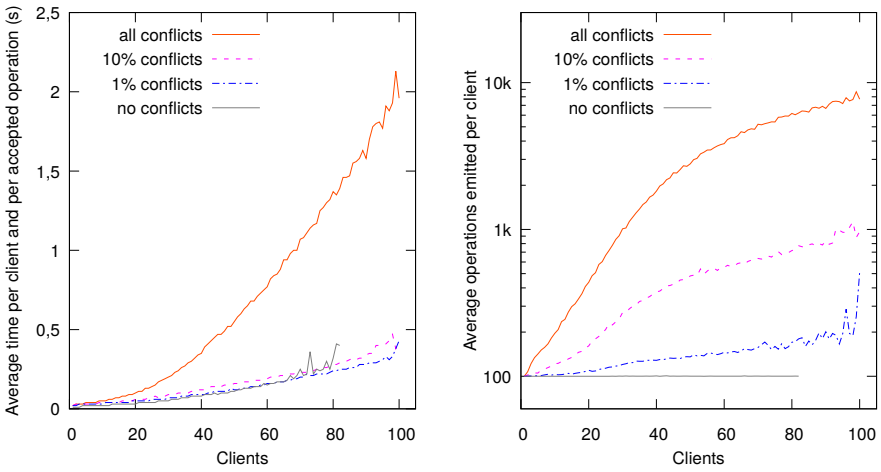


Fig. 5. Time (left) and operations (right, logscale on y-axis) for handshaking 100 operations per client under several scenarios. Parallel clients and server are simulated on a same laptop computer (8 cores, 2.60 GHz, 16GB RAM).

4.2 Evaluation

We simulated a server and many clients performing either commuting or conflicting operations. In the *no conflicts* scenario, all clients move each a different label. In the *all conflicts* scenario, all clients perform conflicting operations on a same label. Indermediate scenarios are when clients update one out of 100 or 10 random attributes of a label (*1% and 10% conflicts*). Parallel clients attempt operations every 0.01 second. When a client receives a DENY, he tries another operation, until 100 of his operations have been accepted. Such a stress test is not realistic but enable to see how robust the protocol is.

Figure 5 shows that the best-case scenarios behave well with up to 100 simultaneous clients. The clients can update labels even without having received previous operations: Starting from 3 clients, more than 99% of the (commutative) operations are here delayed. The worst-case scenario handles until about 20 clients simultanously emitting conflicting operations. Above this number, the server slows down but is stil not stalled.

Finally, we performed client-side evaluation, with until 10 people working on the actual Dezrann clients and annotating collaboratively a score. No server stall was recorded, even when one sometimes feel that an operation was denied.

5 Conclusion, Availibility and Perspectives

We proposed a protocol based on Operation Commutations to handle simultaneous collaborative editions on labels on a music score. This simple protocol is between OT and lock mechanisms. Conflicts do occur but may be ef-

ficiently handled. Simulations show that it enables a scalable use by multiple clients, in scenarios both without and with conflict, and human evaluation confirms that the platform works for a few simultaneous people, answering the needs for the envisaged use cases. Prototype implementation is available in the `dez-{server,client}/dez-collab` directories on `gitlab.dezrann.net`.

Perspectives on the protocol includes improving client (re)joining a collaborative channel, notably by better handling and compressing history to avoid spurious transfers of the full *history*[ ] table and also by using the local storage of the web application. The protocol could also use more OT techniques for text edition in tags and comments. On the development part, perspectives include improving the user interfaces to make the fully useable by everyone, using authentification mechanisms to grant collaborative edition accesses, and finally conducting more user tests both with adults and children groups.

References

1. J. T. Allison et al., NEXUS: Collaborative performance for the masses, handling instrument interface distribution through the web. In *New Interfaces for Musical Expression (NIME 2013)*, 2013.
2. G. Bagan et al., Modélisation et visualisation de schémas d'analyse musicale avec music21. In *Journées d'Informatique Musicale (JIM 2015)*, 2015.
3. P. Couprie. iAnalyse : un logiciel d'aide à l'analyse musicale. In *Journées d'Informatique Musicale (JIM 2008)*, 2008.
4. C. A. Ellis and C. Sun. Operational transformation in real-time group editors: issues, algorithms, and achievements. In *ACM Computer supported cooperative work*, 1998.
5. M. Giraud. Using Dezrann in musicology and MIR research. In *Digital Libraries for Musicology (DLfM 2018)*, 2018.
6. M. Giraud et al., Dezrann, a web framework to share music analysis. In *Int. Conf. on Techn. for Music Notation and Representation (TENOR 2018)*, 2018.
7. H. H. Hoos et al., The GUIDO music notation format. In *Int. Computer Music Conf. (ICMC 1998)*, 1998.
8. D. Huron. Music information processing using the Humdrum toolkit: Concepts, examples, and lessons. *Computer Music Journal*, 26(2):11–26, 2002.
9. A. Imine et al., Proving correctness of transformation. In *Eur. Conf. on Computer-Supported Cooperative Work*, 2003.
10. G. Lepetit-Aimon et al., INScore expressions to compose symbolic scores. In *Int. Conf. on Techn. for Music Notation and Representation (TENOR 2016)*, 2016.
11. J. Malloch et al., A network-based framework for collaborative development and performance of digital musical instruments. In *Int. Symp. on Computer Music Modeling and Retrieval (CMMR 2007)*, 2007.
12. E. Paris et al., KIWI : Vers un environnement de création musicale temps réel collaboratif. In *Journées d'Informatique Musicale 2017* (JIM 2017), 2017.
13. L. Pugin et al., Verovio: A library for engraving MEI music notation into svg. In *Int. Society for Music Information Retrieval Conf. (ISMIR 2014)*, 2014.
14. C. S. Sapp. Verovio Humdrum Viewer. In *Music Encoding Conf. (MEC 2017)*, 2017.
15. M. Shapiro et al., Conflict-free replicated data types. In *Symp. on Self-Stabilizing Systems*, 2011.

Distributed Scores and Audio on Mobile Devices in the Music for a Multidisciplinary Performance

Pedro Louzeiro[1] *

Universidade de Évora, Évora, Portugal
Centro de Estudos de Sociologia e Estética Musical (CESEM), Lisbon, Portugal
pedrolouzeiro@gmail.com

Abstract. In an attempt to uncover the strengths and limitations of web technologies for sound and music notation applications, driven by aesthetic goals and prompted by the lack of logistic means, the author has developed a system for animated scores and sound diffusion using browser-enabled mobile devices, controlled by a host computer running Max and a web server. Ease of deployment was seen as a desirable feature in comparison to native application computer-based systems – such as Comprovisador, a system which has lent many features to the one proposed herein. Weaknesses were identified motivating the design of mitigation and adaptation strategies at the technical and the compositional levels, respectively. The creation of music for a multidisciplinary performance has served as a case study to assess the effectiveness of those strategies.

Keywords: Animated Notation, Electro-acoustic Composition, Web Applications, Multidisciplinary Performance, Networked Music Performance

1 Introduction

Recently, we were invited to compose music for a multidisciplinary performance entitled GarB'urlesco, which includes elements of theatre, dance, costume design and music. In addition to the new music created for this purpose, the performance also includes pieces of baroque music, played on period instruments. Since the theme evokes the sociocultural contrasts that can be observed in the Algarve – a region marked by the seasonality of tourism – it seemed aesthetically relevant to include electro-acoustic elements in the composition. However, there were no logistical or financial resources for this. Not allowing ourselves to give up, we looked for alternatives in recent examples in the field of network musical performance where mobile devices are used as musical instruments, sound projectors, animated scores or as an interface for audience participation.

Regarding our previous experience, as of 2015 we have been developing Comprovisador, a system designed to enable mediated soloist-ensemble interaction

* I should like to thank my supervisors Christopher Bochmann and António de Sousa Dias for their advice and Sara Ross for proofreading. I would also like to thank Elsa Santos Mathei and all artists who made GarB'urlesco possible.

using machine listening, algorithmic compositional procedures and dynamic notation, in a networked environment [14, 15, 13]. In real-time, as a soloist improvises, Comprovisador's algorithms produce a score that is immediately sight-read by an ensemble of musicians, creating a coordinated response to the improvisation. To this date, Comprovisador has been used in ten public performances. In each performance, the system is presented with new features. The introduction of composed electronics in synchronisation with the animated scores is to be expected soon.

Like any other tool, Comprovisador has its strengths and its weaknesses. Weaknesses are often problems for which a solution has not yet been found or represent a trade-off from other aspects considered more important. One such case is the fact that Comprovisador is not compatible with mobile devices (namely, tablets). Not only does it require proper computers but it also requires software installation procedures [1] that are not very complex for someone experienced, but can be so for someone who is not computer oriented. Thus, it can be tedious or discouraging for users and it always entails increased preparation time when preparing a concert. These issues would be problematic in the context of GarB'urlesco and would not solve the fundamental issue of sound projection logistics.

For Comprovisador's objectives, processing speed and reliability of laptop computers running native applications outweigh ease of deployment of tablets running web applications. Other applications with different objectives may benefit more from a mobile device-based approach. Apart from musical performance with real-time generated scores, we envisage possibilities in the educational field – Soundslice [10], a web platform where users can learn to play pieces of music through dynamic notation synchronised with Youtube videos is a good example, but there are also possibilities in the fields of ear training and sight reading (see [16]).

In the field of music and multimedia networked performance, there are many examples that use mobile devices with as many different strategies. Following are only a few examples. "Flock" [8, 9], a piece by Jason Freeman for saxophone quartet, video, electronic sound, dancers, and audience participation premiered in 2007, uses PDA's mounted on each player's instrument. Those devices display music notation generated from the locations of musicians, dancers, and audience members as they move and interact with each other. Decibel ScorePlayer [11] is an iPad application developed by the Decibel New Music Ensemble, a group led by composer Cat Hope. This application enables network-synchronised scrolling of proportional colour music scores and audio playback. It has been used by many composers worldwide. Cheng Lee proposes an approach for incorporating computer music and virtual reality practices into a multimedia performance installation requiring the audience members to use their own smartphones as 360-degree viewing devices [12]. The author also proposes the use of wireless speakers carried around the venue as a means of achieving immersive sound and

[1] Both host and client applications of Comprovisador run in the Max environment [18] using the bach library [4] and also Java.

music effects in substitution for a multi-channel surround-sound system. Composer Jonathan Bell, who also has used bluetooth speakers as networked sound projectors, has created a system called SmartVox [5] – a web-based distributed media player as notation tool for choral practices. The audio-visual scores are created with the bach environment for Max. In a choral context, singers hear (using earphones) and see their own part displayed in the browser of their smartphone. The whole is synchronised through the distributed state of the web application. Of the examples given here, only SmartVox uses purely web technologies. Finally, we should mention a.bel [6] – a system presented in 2015 at Casa da Música, in Porto, Portugal, in a concert where almost 1000 smartphones were used as musical instruments by the members of the audience. The event featured pieces by four composers – Carlos Guedes, José Alberto Gomes, Neil Leonard and Rui Penha – using different approaches to audience participation via their smartphones and the a.bel system.

In this paper, based on a case study – the music composed for GarB'urlesco, we will attempt to demonstrate a possible application for web resources (HTML5, JavaScript (JS) and open-source libraries) which includes animated precomposed scores and distributed sound diffusion on a local network. Since synchronisation is not easily achievable in this context, we will discuss certain mitigation and adaptation strategies that were adopted, regarding the technical and the compositional sides, respectively. The compositional strategies also took into account the sound characteristics of this type of devices.

The motivation for carrying out this practical application is therefore related to research on easily deployable solutions and sharpened by the will to perform electro-acoustic music in a context with scarce logistical resources.

2 HTML5-based Solutions

General advantages of mobile devices for animated score and sound applications include ubiquity (most especially in the case of smartphones), ease of transportation and set-up (from one's pocket directly to the music stand), and a fair processing power despite being small and lightweight. As for the advantages of web-based software we count being completely cross-platform and cross-type (laptops, tablets, smartphones, etc.), having no need for additional software installation (any required libraries are loaded by the browser at runtime), being free of charge, capable of performing animation at an adequate frame rate (HTML5 Canvas element / `window.requestAnimationFrame()` method, which we find to have a superior performance than Max's `jsui` or `lcd` options (cf. [15])), and including access to powerful open-source tools for graphics and sound (p5.js, p5.sound [1], tone.js [2]) and for interfacing with Max (Miraweb and its underlying library – xebra.js [3]).

This approach also has disadvantages, specifically the inability to use UDP [2] and the unavailability to the best of our knowledge of an open-source [3] music notation library with suitable quality for generative applications. Furthermore, there are important timing issues with three different origins: 1) network latency, which is aggravated by the use of the Transmission Control Protocol (TCP) (since UDP as an alternative is unavailable); 2) imprecise JavaScript clock (accessing the audio subsystem's hardware clock through the Web Audio API can improve precision although not to our desired levels, as will be discussed in Section 3.1); and 3) latency originating in other factors.

Considering the above disadvantages, can we nonetheless accomplish interesting musical results by implementing mitigation and adaptation strategies? This question is what we attempt to answer in the following sections.

3 Case Study

GarB'urlesco is a multidisciplinary performance with elements of theatre, dance, costume design and music (both early music and purposely composed music), using period instruments and a traditional cane flute, composed electronics for networked mobile devices and animated score. The score uses standard notation which is dynamically updated also featuring a bouncing ball cueing system for synchronisation between instrumentalists and electronics.

Regarding space, the devices are arranged according to Fig. 1 (left). Smartphones are hidden under the audience's chairs, which are arranged in double lines on each side of the room. Device number 1 (tablet) is the only device in charge of displaying the score and it sits on the flautist's music stand.

The intent of hiding the devices in close proximity to the audience is to create immersive sound effects while causing some strangeness (hearing sound but not being able to see its source) and to compensate for the weak sound output of these devices.

In addition, there is a host computer running Max that coordinates sound and score events, being also connected to a small 2.0 sound system, located in the musicians' space, enabling low frequency sound effects.

We have designed a system that uses xebra.js to establish WebSocket [4] communication between Max and the browser of each device. Besides Max, the host computer runs a server hosting two HTML files, two JavaScript libraries (xebra.js and tone.js), various audio files and the score (comprised of png files, designed in Max using the `bach` library [4] for notation and Computer Assisted Composition (CAC) features – see Section 3.1). One of the HTML files is destined for audio playback only and the other for displaying the score as well (device number 1).

[2] UDP (User Datagram Protocol) is a communications protocol used primarily for establishing low-latency and loss-tolerating connections between network applications.

[3] Paid solutions are not included in the scope of this research.

[4] WebSocket is a computer communications protocol, providing full-duplex communication channels over a single TCP connection.

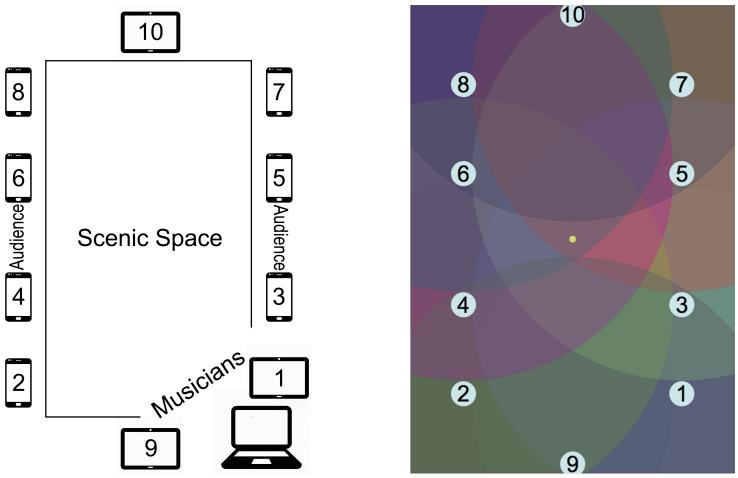


Fig. 1. Left: space arrangement of mobile devices in GarB'urlesco; right: Swarm Spatialiser GUI - `nodes` Max object.

The loading steps occur as follows: each browser is pointed to the appropriate HTML file; the user is prompted to input a unique label for easy identification by Max; the relevant audio files and, if applicable, the score png files are then loaded by the script; meanwhile, a WebSocket connection is established with Max; by user interaction, the audio context is started up and Max is notified of the client's status (online).

3.1 Technical Problems and Mitigation

Battery operated devices turn off their screens and other resources when left idle with no user interaction. In our case, it makes sense to leave the screen off as it saves up battery power, avoiding the need to plug everything to the mains, also helping conceal the devices. The problem is that other resources such as sound output and Wi-Fi connection also go to sleep or run in low consumption mode, thus hampering performance.

One possible workaround for this issue is to play a dummy sound file every few seconds, triggered from Max over the network. This keeps both resources awake [5]. The dummy sound file must contain some audio data otherwise it is dismissed by the system. On the other hand, we obviously could not do with an audible repeating sound. Hence our sound file contains a very short (100ms), low amplitude (-21dB) and low frequency (30Hz) sine wave which cannot be heard from the devices' speakers though having enough information to trick the system into keeping both resources active.

In Comprovisador, we had implemented a system to attain suitable synchronisation which was based on Dannenberg's concept of "time-flow" [7] – specifi-

[5] – in devices running older Android versions – see Section 4.

cally, level 1 synchronisation [6] which uses time-stamps. For time-stamps to have actual meaning within a network, all machines must agree on the time. This problem can be addressed by using a clock synchronisation approach similar to the Network Time Protocol (NTP).

NTP is based on low-latency UDP. However, browsers do not allow the use of this protocol for security reasons, which raises problems for applications like ours.

After running a few tests with the WebSocket protocol we have realized that, in a local network, the round-trip time values were very unpredictable, ranging between below one millisecond and above two hundred milliseconds. Hence, we have developed an algorithm that queries the server time repeatedly, during 2 seconds. By probability, the fastest round trip within this interval will be under eighty milliseconds – although very often under one millisecond. By selecting this iteration we maximise precision and use the reported server time to make the necessary adjustments to the client's clock. The algorithm is presented below.

```
function queryNow(){    // Called repeatedly during a 2s interval
  if (querying) {
    var wrong_time = now(); // client time is assumed wrong
    var message = ["queryNow",wrong_time];
    sendMax(message);       // query Max to get server time
  }
}
function timedResponse(t){  // Called upon Max response
  var wrong_time = t[0];    // Max echoes the client time
  var reported_time = t[1]; // and reports server time
  var received_time = now(); // Client time-stamps the response
  var round_trip = received_time - wrong_time;
  if (round_trip < best_lap) {
    best_lap = round_trip;  // Store the fastest iteration
                            // calculate our offset
    ntp_off = reported_time - round_trip/2 - wrong_time;
  }                         // after 10ms, recall queryNow
  Tone.context.setTimeout(queryNow, 0.01);
}
function queryEnd() {       // Called when the 2s interval ends
  querying = false;
  Tone.Transport.seconds += ntp_off; // adjust client time
}
```

[6] This concept aims at solving synchronisation problems in real-time music and media systems. The author describes four approaches to synchronisation in increasing levels of sophistication: Synchronisation Levels 0 through 3. Level 1 consists on applying time-stamps to events, computing the events in advance within a "control stage" and delivering the computed events with time-stamps to a "rendering stage". There, events are delayed according to time-stamps in order to produce accurately timed output.

On the server side, time is obtained with the Max `cpuclock` object, which is the most accurate [18]. On the client side, we obtain it with `Tone.Transport` (from the tone.js library) which is based on Web Audio's `audioContext.currentTime` property. It is indeed more accurate than the JS clock. Nevertheless, it is not as accurate as we would like it to be. A simple experiment consisting on scheduling a repeated event every 500ms with `Tone.Transport.scheduleRepeat` using various latency hint values revealed noticeable jitter in every device we have tested (Android only).

We have encountered additional latency originating in other factors, at a later stage than the network and the Audio Web clock. We believe this is originated in the audio/video hardware but cannot be sure, since it is outside our area of expertise. The latency value is different from device to device but, within the same device, it stays relatively stable.

To tackle this problem we have implemented a graphic user interface (GUI) allowing manual adjustment of the latency compensation for audio and also for video (score version).

As for the score, since we were unable to find a suitable JS notation library, we have turned to the idea of using image files. This is possible thanks to an image-export feature included in the forthcoming version 0.8.1 of bach, on which we were allowed to beta-test a pre-release.

Our approach is well suited for precomposed music and can, in the future, be modified for real-time generative scores. Here are its characteristics: measures are unitary (1/4, 3/8 or 3/16); each measure consists of an independent png file of a constant pixel size; the number of measures per system depends on screen resolution and is optimized for 8 measures on the more common screens; measures are dynamically updated in a cycle in such a way as not to disturb the reading process – replacing the ones that have been read at the distance of half a system (half a cycle). This approach was adapted from a previous one recently introduced in Comprovisador, the difference being that, in Comprovisador, notation is generated and drawn in real-time using bach objects.

Another approach inherited from Comprovisador is the bouncing ball as a visual synchronisation and score navigation device. Other authors have used similar approaches (see [19, 21]) and consider them preferable when comparing to other types of score navigation strategies [17].

In our web-based application, two overlaid canvas elements are used: one holds the score while the other renders the bouncing ball. The former is redrawn the least number of times possible – once only whenever the measure is updated and only in the corresponding rectangle. The latter is redrawn around 60 times per second (with `requestAnimationFrame()`). It would be possible to use only one canvas but it seemed unnecessary to redraw all the png images in every frame.

Since rendering of the bouncing ball is controlled by a different clock (JS instead of the Audio Web clock – cf. [20]) a drift may occur. The correction for such drift is obtained as follows:

– Max host sends the message `syncScore(n, t)` every few measures, where `n` is the measure to reach at time `t`;
– if at `t` the bouncing ball is lagging behind, then jump forward to `n`;
– else if ball is early, reduce velocity by a factor in order to reach `n+1` on time `t+period` (skew).

The last step of this algorithm helps avoiding jitter, which would occur if the bouncing ball had to jump backward.

3.2 Compositional Problems and Adaptation

Several problems are faced when composing mixed music with mobile devices as sound projectors:

low power output – a smartphone cannot balance with an acoustic instrument;
poor quality – absence of low frequency spectrum and some degree of distortion;
latency – even with our mitigation strategies, it is not possible to have the required level of timing precision to perform accurate spatialisation.

Regarding the lack of power, we adapt by using number – at least one device for every four people – and textural reduction, using electronics mainly against solo instruments. On the issue of poor quality, we can restrict to using high-pitched sounds. As for latency, "if you can't beat it, join it", which in this case means to embrace non-simultaneity – to use granulation-based effects.

In fact, we have built a granular synthesiser controlled from a bach score. Thanks to bach's slot system, it allows sequencing all the granular synth's temporal parameters, as well as the transposition parameters (the musical notes) in the same notation environment used for the instrumental parts (see Fig. 2).

Although accurate spatialisation is not feasible, we have conceived a strategy thus allowing to convey spatial sensations encompassing temporal inaccuracies. It uses the characteristics of granulation – namely, the probability of a given grain to be emitted by a particular source.

In Fig. 1 (right), we see a GUI made with the object `nodes`. Here, each ellipse represents the field of influence of each source while the cursor's position in relation to each ellipse represents the probability weight of a grain, at a given moment, to be emitted by the respective source.

To avoid phase problems due to temporal inaccuracies, a grain can never be emitted by more than one source simultaneously. The amplitude is based on the same weight assignment GUI but with normalised and scaled values.

This system is intended to simulate, to a certain extent, a graphic system of particles where each particle revolves around a point in space, with apparent free will. Although we never see the centre point, we get a sense of where it is by the way particles express their desire towards it. The same happens in our granular spatialiser – which we name Swarm Spatialiser.

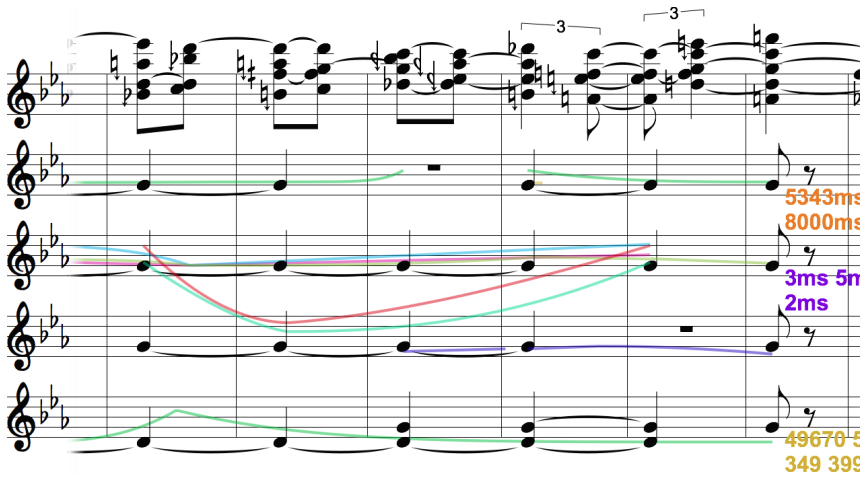


Fig. 2. A `bach.score` object featuring several slots (types *function* and *intlist*).

However, when we apply an automated trajectory generator to our GUI to get a stream of probability weights, and then apply further calculations of norm and scale to those values before feeding them to the granular synth, our system's CPU utilization becomes dangerously high. We have solved this by recording the stream of weights (post-calculations) to a `bach.roll` object as slot content of type llll [7]. Then, when we read back the values from `bach.roll` as a lookup table, we avoid performing all the calculations within the trajectory generator (Lissajous curves and other functions), the `nodes` GUI (translation of coordinates into probability weights) and beyond.

Another advantage of using bach is its microtonal support which we have taken advantage of (see Fig. 2). In fact, some traditional melodies were used as a basis for the composition and those melodies contain various microtonal inflections. Therefore, it seemed relevant to extrapolate that feature to the remaining musical elements. Bach works in midicents, which our granular synth accepts. Regarding notation, it supports pitch breakpoints in the duration lines, allowing glissandi to be defined and played back.

4 Results

The system was assessed by composer and musicians during rehearsals (April to July 2019) and performance (6 and 7 July 2019).

The animated score was found to perform suitably, with faster performance of the bouncing ball when compared to previous experiments in the Max environment. Also, the appearance of the staff, despite being comprised of individual image files, is seamless.

[7] llll stands for Lisp Like Linked List.

Synchronisation between flautist and electronics guided by the bouncing ball was considered effective, but it must be taken into account that there were no events requiring very precise synchronisation (as part of our adaptation strategy). Other instruments' lines were subject to that of the flute and were easy to follow along with the paper version of the score.

Manual latency compensation was found relatively easy to adjust sensorially. Once adjusted, there was some probability of a change in the latency value, causing the device to be slightly out of sync. In slower devices running Android, this was more noticeable. However, this probability was not very significant. Moreover, our compositional strategy accounts for these imprecisions.

The dummy sound workaround for keeping resources active was effective with many of the devices available to us. However, it did not work with iOS devices nor with devices running Android version 8 or later. On iOS devices, the solution was to simply set the display to not enter sleep mode. Recent Android devices revealed more disadvantages: developer options had to be enabled and devices had to be connected to a power source in order to stay awake during the show.

At the sequencing level, the probability-based granular spatialiser was considered effective, being analogous to a graphic particles system simulating swarm intelligence. The use of the `bach.roll` for trajectory sequencing has enabled better performance by eliminating the need to perform a substantial number of calculations, replacing it with a lookup table approach.

Furthermore, using bach to sequence our granular synth's pitch and temporal parameters has proved advantageous in regards to the integration with the instrumental melody representation. It is noteworthy that the slot system enables a kind of interaction similar to the one enabled in Digital Audio Workstations (DAW) for automation control, with the advantage of a deeper integration with notation. On top of that, it allows us to sequence data for instruments created in Max which therefore are not available in any DAW software.

Finally, at the composition level, our adaptation strategies have made it possible to circumvent inherent problems in this type of system and achieve results with pleasing aesthetics. Regarding the sound spectrum restriction and the instrumental texture reduction, our choice was to use granulated instrumental sounds taken from the ensemble in order to create a dialogue with the respective acoustic instruments, as a duet. This approach offered timbre cohesion and balance, overcoming the poor quality and sound power of the devices. With regards to sound design, the system adapted well to the effects required by the narrative – namely, 'countryside' (represented by the bells of a flock of sheep), 'sea' (filtered white noise and seagulls), 'frantic tourism and its bipolar seasonality': extreme calmness vs chaos (smartphone ringtones, camera shutters, DTMF tones [8] and glitchy sounds). These sounds were chosen firstly based on their energy in the high spectrum and secondly because they were in some way associated with the concept of flock – sheep, seagulls, tourists with their smartphones and cameras –

[8] Dual-tone multi-frequency (DTMF) signaling is a telecommunication signaling system using the voice-frequency band over telephone lines between telephone equipment and other communications devices and switching centers.

and therefore appropriate for our Swarm Spatialiser. Sensations of sound source displacement were felt clearly around the room as expressed by a few audience members who came to us once the show ended.

Also appreciated was the satirical use of the ringtones emitted by actual smartphones: it starts with only one phone ringing, thus leading the audience to think someone forgot to turn off their phone. After a short while and a blatant ringtone crescendo, the purpose becomes obvious.

In addition to the mobile devices, it was possible to obtain sound effects with stronger dynamics and lower frequencies by using a 2.0 sound system. Effects consisted in certain soundscapes and reinforcement of instrumental passages.

GarB'urlesco was conceived for a particular venue, in Lagos, Portugal. If the possibility ever comes of performing it elsewhere, some adaptation will likely be needed. For example, a larger venue will require a larger number of devices – which could raise problems regarding network capability – or the use of portable speakers, which would change the original sound characteristics.

5 Conclusions

A system of this nature can hardly compete with native application computer-based systems for distributed music in contexts where the music requires great accuracy and reliability. A notation system like Comprovisador takes advantage of the timing accuracy of Max's scheduler and bach's CAC tools enabling distributed computing of real-time generated scores with great flexibility and scalability. An array of speakers will deliver the full range of audible frequencies with the dynamic range of an orchestra and unsurpassable timing precision. But these systems involve significant preparation time and/or logistic resources.

Hence, in situations where it is possible to adapt the music creation to the idiosyncrasies of a web-based system, it is possible to take advantage of its best features (notably, ease of deployment) and achieve aesthetically pleasing results.

The case study presented herein showed the usefulness of some of the features available in the bach library for achieving the desired results. On one hand, the new image-export feature allowed us to create a dynamic, nice-looking score on the browser using png files as discrete measures. On the other hand, the slot system present in the `bach.score` and `bach.roll` objects allowed integrated sequencing of parameter values and microtonal notation for our granular synth and respective Swarm Spatialiser. These latter tools were crucial in the accomplishment of the outlined compositional strategies.

References

1. p5.js, `https://p5js.org`, Processing Foundation
2. tone.js, `https://tonejs.github.io`
3. xebra.js (2018), `https://cycling74.github.io/xebra.js/index.html`, cycling74
4. Agostini, A., Ghisi, D.: A max library for musical notation and computer-aided composition. Computer Music Journal 39(2), 11–27 (2015)

5. Bell, J.: Audiovisual scores and parts synchronized over the web. In: Bhagwati, S., Bresson, J. (eds.) Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18. pp. 17–23. Concordia University, Montreal, Canada (2018)
6. Clément, A.R., Ribeiro, F., Rodrigues, R., Penha, R.: Bridging the gap between performers and the audience using networked smartphones: the a.bel system. In: Proceedings of the International Conference on Live Interfaces (2016)
7. Dannenberg, R.B.: Time-flow concepts and architectures for music and media synchronization. In: Proceedings of the 43rd International Computer Music Conference. pp. 104–109. Shanghai (2017)
8. Freeman, J.: Extreme sight-reading, mediated expression, and audience participation: Real-time music notation in live performance. Computer Music Journal 32(3), 25–41
9. Freeman, J.: Flock (2007), http://distributedmusic.gatech.edu/flock/
10. Holovaty, A., Richardson, C., O'Riordan, E.: Soundslice (2019), https://www.soundslice.com
11. Hope, C., Vickery, L.: The decibel scoreplayer – digital tool for reading graphic notation. In: International Conference on Technologies for Music Notation and Representation. Paris (2015)
12. Lee, C.: Multimedia performance installation with virtual reality. In: Proceedings of the 43rd International Computer Music Conference. pp. 347–350. Shanghai (2017)
13. Louzeiro, P.: Mediating a comprovisation performance: the Comprovisador's control interface. In: Proceedings of the 43rd International Computer Music Conference. pp. 362–367. Shanghai (2017)
14. Louzeiro, P.: Real-time compositional procedures for mediated soloist-ensemble interaction: The Comprovisador. In: Agustín-Aquino, O.A., Lluis-Puebla, E., Montiel, M. (eds.) Mathematics and Computation in Music: 6th International Conference, MCM 2017, Mexico City, Mexico, pp. 117–131. Springer International Publishing, Cham (2017), https://doi.org/10.1007/978-3-319-71827-9_10
15. Louzeiro, P.: The Comprovisador's real-time notation interface (extended version). In: Aramaki, M., Davies, M.E.P., Kronland-Martinet, R., Ystad, S. (eds.) Music Technology with Swing, pp. 489–508. Springer International Publishing, Cham (2018), https://doi.org/10.1007/978-3-030-01692-0_33
16. Louzeiro, P.: Improving sight-reading skills through dynamic notation – the case of Comprovisador. In: Bhagwati, S., Bresson, J. (eds.) Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18. pp. 55–61. Concordia University, Montreal, Canada (2018)
17. Picking, R.: Reading music from screens vs paper. Behaviour & Information Technology 16(2), 72–78 (1997)
18. Puckette, M., Zicarelli, D., Sussman, R., Clayton, J.K., Bernstein, J., Nevile, B., Place, T., Grosse, D., Dudas, R., Jourdan, E., Lee, M., Schabtach, A.: Max 7: Documentation, https://docs.cycling74.com/max7/
19. Shafer, S.: VizScore: An on-screen notation delivery system for live performance. In: Proceedings of the International Computer Music Conference. pp. 142–145. Denton, TX (2015)
20. Wilson, C.: A tale of two clocks - scheduling web audio with precision (2013), https://www.html5rocks.com/en/tutorials/audio/scheduling/
21. Zagorac, S., Alessandrini, P.: ZScore: A distributed system for integrated mixed music composition and performance. In: Bhagwati, S., Bresson, J. (eds.) Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18. pp. 62–70. Concordia University, Montreal, Canada (2018)

The BabelBox: an Embedded system for Score Distribution on Raspberry Pi with INScore, SmartVox and BabelScores

Jonathan Bell[1], Dominique Fober[2], Daniel Fígols-Cuevas[3], and Pedro Garcia-Velasquez[4]

[1] CNRS, PRISM "Perception Representations Image Sound Music",
[2] GRAME CNCM
[3] King's College London
[4] BabelScores
belljonathan50@gmail.com fober@grame.fr
daniel.figols-cuevas@kcl.ac.uk pgarcia@babelscores.com

Abstract. The slow but steady shift away from printed text into digital media has not yet modified the working habits of chamber music practitioners. If most instrumentalists still heavily rely on printed scores, audiences increasingly access notated music online, with printed scores synced to an audio recording on youtube for instance. This paper proposes to guide the listener and/or the performer with a cursor scrolling on the page with INScore, in order to examine the consequences of representing time in this way as opposed to traditional bars and beats notation. In addition to its *score following* interest for pedagogy and analysis, the networking possibilities of today's ubiquitous technologies reveal interesting potentials for works in which the presence of a conductor is required for synchronization between performers and/or with fixed media (film or tape). A *Raspberry Pi*-embedded prototype for animated/distributed notation is presented here as a *score player* (such as the *Decibel ScorePlayer*, or *SmartVox*), in order to send and synchronize mp4 scores to any browser capable device connected to the same WIFI network. The corpus will concern pieces edited at BabelScores, an online library for contemporary classical music. The BabelScores pdf works, composed in standard engraving softwares, will be animated using INScore and video editors, in order to find strategies for animation or dynamic display of the unfolding of time, originally represented statically on the page.

Keywords: Distributed scores, animated notation, music publishing, embedded system, networked music performance

1 From paper towards animated notation: the quantum leap

A great majority of orchestral musicians today would still consider cumbersome the idea of replacing sheet music by screens. And yet, as early as 1999, thorough research had already been undertaken (by J. McPherson [24]) in a realm which is now identified as *animated notation* or *screen-scores*.[5] In his works and its surroundings [25], questions

[5] The term *screen-scores* is attributed to Lindsey Vickery and Cat Hope. [10]

relative to scrolling vs turning pages, resolution and number of systems per pages, were addressed for the first time, and already gave substantial elements of response to issues which now concern most music theory educational games.

1.1 Where gaming meet musical notation

The representation of time in the age of *screen scores* opens a wide array of questions. Thinking of tablature[6] notation, if we understand scores as a set of prescribed actions to be performed in time, video games such as *Guitar Hero* or *Beatmania* will show that such new forms of music-making offer unprecedented control over the sequential realisation of rhythmic patterns. If, on the other hand, notation is defined as something to be interpreted, and performed on an acoustic instrument, perhaps these games will hardly be recognised as "musical" experiences by musicians.

Compared to traditional score written on paper or *Common Music Notation* (hereafter referred to as CMN), graphical notation scrolling on a screen offers a more mimetic (more direct, less codified) approach to music reading, which evokes video games in many ways. With the exception of the Decibel Score Player (see Mezareon for instance) or composers cited in the animatednotation.com website, this rather ludic or *gamified* approach to score reading has only gained visibility in education so far. In the pedagogical context indeed, recent applications such as *Simply Piano* take great advantage of their ability to validate/invalidate the performer's actions, again as a way to *gamify* apprenticeship of music reading. Unfortunately *Simply Piano* is an exception, and the general tendency in software development (as shown in the case of *Guitar Hero, Synthesia* or *Beatmania* for instance) is simply to bypass CMN altogether (both for in term of pitches and rhythm). Whilst some of this games might help developing music skills among amateur practitioners (such as those youtube videos tagged [Piano Tutorial] (Synthesia)), effort remains to be made in order to find bridges between those emerging amateur practices and the art of musical notation as it is used by contemporary classical composers and performers, both still active in concert halls and academia.

1.2 Animation in the context of Distributed Musical Notation

Ryan Ross Smith's ***animatednotation.com*** website features many examples of composers elaborating scores taking advantage of the possibilities of screen scores. Most of these example rely on the projection on one single score projected on a screen visible to the audience and the performers. The networking capacities of today ubiquitous technologies would however easily allow each performer to receive only his own part

[6] Tablature, as opposed to common music notation, can be conceived as prescriptive notation, in the sense defined by Mieko Kanno: 'Prescriptive notation specifies the means of execution rather than the resultant configurations of pitch and rhythm' (Kanno, 2007, p.1). The distinction between prescriptive and descriptive notation (or common music) was already discussed in the fifties, as can attest the following statement by Charles Seeger: 'Prescriptive and descriptive uses of music writing, which is to say, between a blue-print of how a specific piece of music shall be made to sound and a report of how a specific performance of it actually did sound (...)' (Seeger, 1958, p. 1).

of the score. Indeed, composers and researchers increasingly acknowledge the strong analogy which can be drawn between the traditional 'score and parts' musical practice led by a conductor, and the modern distributed systems or web applications (Zscore [22] - MASD [14] - SmartVox [2]), in which multiple clients coordinate their actions by passing messages to one another. Several attendees of the Tenor Conference[7] have proposed elements of response in an emerging realm which can be called "distributed musical notation". Some performance-oriented systems (INScore[17], SmartVox [2] [3], Zscore [22], Decibel [5] [6], MaxScore [7], Comprovisador [23]) endeavour to distribute and synchronise each part of the score on the performer's devices (whether Smartphones, tablets or laptops).

1.3 Animated notation and the composers of the new complexity

Only a few composers of contemporary classical notated music feel the necessity to acknowledge this sudden growth of animated/distributed notation, since the software they use (Finale, Sibelius or pen and paper in great majority) are designed to render still images. Also these scores are dedicated to classically trained performers, all familiar with CMN and in great majority also aware of complex experimental forms of "static" notation. In academia, composers and music analysts are trained to read and follow these complex scores in which the notation is sometimes overloaded and detached from the acoustic result, as can be exemplified with the experimental music of the so called *new complexity*, with extremely complex rhythms (famously led by Brian Ferneyhough), or with some of its more recent (post-Lachenmanian) manifestations with large amounts of extended techniques and graphical notation.

The *Score Follower* project helps such composers getting their music heard and understood by simply synchronising a recording to each page of the score and share it on social media (youtube). To some unexperienced readers/listeners, the sound to sign relationship may still be difficult to follow which is one of the reasons why the cursor was introduced here. The representation of musical time in the examples below will propose an attempt of *hybridised* situations in which (often complex) contemporary classical scores take advantage of the possibilities of animated notation (DENM being a major influence [15]). The case of *New Complexity* or post-Lachenman types of aesthetics will be of particular interest here (see Malaussena), as the proposed cursor solution may provide elements of response to composers whose rhythmic complexity seeks *rubato* in the first place. Indeed, beyond the exact realisation of nested tuplets, some of these composers seek in rhythmic complexity the absence of a clear sense of pulse and fluidity.

1.4 Scrolling versus beating time

In the above-mentioned video games and musical examples, wether notation is scrolling from right to left (as in *Simple Piano* or *Decibel ScorePlayer*) or top to bottom (Guitar Hero - [Piano Tutorial] (Synthesia)), the basic principal relies on a continuous or scrolling movement (either of a cursor, or of the score itself) representing the passage of time. Cursors, as will be seen below, present great possibilities of synchronisation, in terms

[7] http://www.tenor-conference.org/

of *duration* rather than *rhythm*: when instruments need to synchronise with electronics or video for instance. In terms of pulse however, cursors remain quite approximate in comparison with the arm movement of a conductor dictating a beat. In none-pulsed music therefore, and in spite of the great conducting tradition in chamber music and orchestral works, such scrolling displays (as in the Decibel ScorePlayer or SmartVox) seem a far more straightforward strategy to obtain synchronisation in comparison to the bars and beats 'encoding' (quantified by the composer's choice of bars and beats) and decoding processes (a compromised interpretation by the instrumentalist, between the rhythmic values written on the page and the gestures of the conductor), inherited from a scoring tradition in which a regular meter was assumed.

We therefore propose here a solution for animation of pre-existing scores, with the help of a cursor scrolling on a static page. INScore[17] is in no small part designed for cursor animation, and will be of particular relevance to extend the practice of animated screen scores to a wider community of composers and performers, through BabelScores[8] in particular.

2 BabelScores, SmartVoxINScore, and Presentations

BabelBox, the project envisaged here proposes a collaboration between two existing technologies (INScore and SmartVox) and BabelScores, a publishing company specialised in contemporary classical scores, which will now be presented.

2.1 SmartVox

SmartVox [2][3] is a distributed web application that delivers and synchronizes audiovisual scores in the video mp4 format to the performer's mobile devices, in compositions involving up to 80 simultaneous performers such as in ***Le temps des Nuages***. The ability to synchronise *screen scores* and *audio-scores* [1] through the browser of the performer's phones allows for various kinds of assistance, such as free movement on stage and around the audience, audio guide for singers, simplified synchronisation with tape and/or visuals... Recent developments include the use of head-mounted displays (HMDs) for technology-aided performance, as in the pieces ***In Memoriam Jean-Claude Risset I*** and ***Mit Allen Augen, In Memoriam J.C. Risset II***.[9]

SmartVox was developed in the *SoundWorks* framework.[10] *SoundWorks* provides a set of services – such as synchronization, network messages, distributed states, creation of groups of clients – that aims to solve problems common to distributed and synchronized web applications centered on multimedia rendering. The framework is written in

[8] Babelscores (https://www.babelscores.com/) is an online score database for classical contemporary music, currently actively supporting the SmartVox project: http://1uh2.mj.am/nl2/1uh2/lgi4u.html.

[9] Those three pieces are respectively available at: https://youtu.be/SyFdR2HiF00, https://youtu.be/hQtyu1dcCaI, and https://youtu.be/ET_OBgFWx04.

[10] *SoundWorks* was initiated by the CoSiMa research project funded by the French National Research Agency (ANR) and coordinated by Ircam.

Fig. 1. Singer wearing HMD for technology-aided performance.

Javascript, with a server side based on Node.js. [11]The SmartVox application consists of two web clients, the player and the conductor, that can be executed in any recent web browser on mobile devices (e.g. smartphones, tablets) and laptops. The real-time communication between clients is achieved through the WebSocket protocol [12] The application is typically deployed over a *local area network*, but it may also be used over the internet.[13]

2.2 The 'BabelBox', a Raspberry Pi Hardware Embedded System Solution for Local NMPs - reformulate

In search of a light plug-and-play dedicated system to be sent over the post, the Raspberry Pi quickly appeared as the best option to host SmartVox on an embedded system. Node.js runs on Raspbian, and SmartVox proved to be very stable on a Raspberry Pi 3, so, once installed, the only two steps for a *0-conf* deliverable hardware were:

- Setting up a static address for a dedicated router (e.g. tp-link...).
- Starting SmartVox at boot.

Starting a script at boot can be done on Raspbian with a file containing the following in the etc/systemd/system:

```
[Unit]
Description=My service
[Service]
ExecStart=/home/pi/Desktop/hello.sh
[Install]
WantedBy=multi-user.target
```

With the hello.sh script containing the following to launch the server:

```
#!/bin/bash
```

[11] https://nodejs.org/en
[12] https://www.w3.org/TR/WebSockets/
[13] https://youtu.be/83ub6-Q5oj0

```
cd /home/pi/Desktop/risset
npm run start
exec bash
```

This low-cost system (less than 65 €, for a Raspberry and a router) now allows the sending of ready-to-use scores. Once the system is power-supplied, all the performers need to do is to join the dedicated Wi-Fi, and type the static IP address of the server on their smartphone/tablet (i.e. for the performers: 192.168.0.100:8000, and for the conductor: 192.168.0.100:8000/conductor). In January 2019, the system was rented to the Caen French conservatoire via BabelScores,[14] thus proposing a rental of performing scores (separate parts) of a new kind.

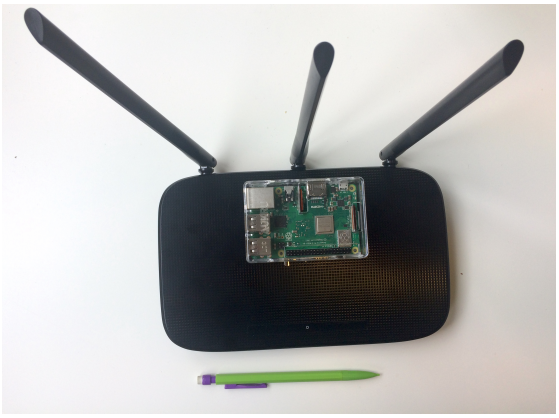


Fig. 2. BabelBox kit.

2.3 BabelScores

BabelScores is a French-based organization arising from the need to circulate and promote contemporary music from the last 50 years, rendering information more easily available to everyone worldwide. Through an international network, BabelScores looks into and selects the works of the most creative, original and innovative composers of the past few decades. BabelScores offers a wide and constantly growing catalogue, setting up a powerful circulation platform especially addressed to instrumentalists, ensembles, orchestras, composers, musicologists, conservatories, universities and festivals throughout the world. BabelScores offers the possibility to consult online all the material in its catalogue. Scores, which are the central element of BabelScores' material, may be consulted bookwise, turning pages, by means of a special reader which allows a comfortable and detailed reading.

[14] Babelscores (https://www.babelscores.com/) currently supports actively supporting the SmartVox project: http://1uh2.mj.am/nl2/1uh2/lgi4u.html. The first piece performed in Caen with the Babelbox is available at the following address : https://youtu.be/wUyw0KQa5Wo

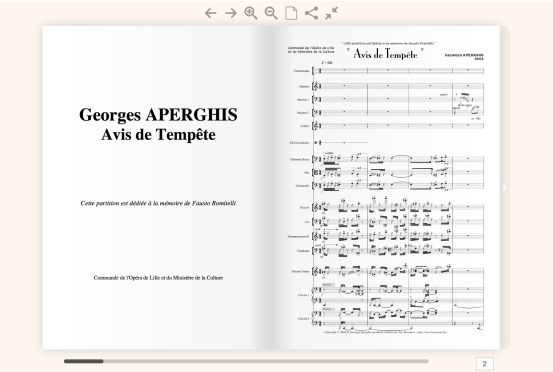


Fig. 3. BabelScore's virtual reader.

In the framework of its purpose to promote, excite, support and diffuse the "written music" creation in connection with web technologies, BabelScores cooperates with traditional institutions as well as with researchers and emerging projects. BabelScores already collaborates with the Bibliothèque Nationale de France (BNF) in order to find and conjointly develop a robust, relevant and pragmatic way to handle music scores in this web and digital era, particularly their preservation and automatic (server-to-server) deposit to archive centres. BabelScores also works on the topic of native digital sketches, in order to asset which formats to use, how to transfer them and how to expose them in the most meaningful way in the BnF's collections once processed. BabelScores collaborate also with researchers and emerging projects such as SmartVox and InScore to supports applied research in the domain of notation using web technologies. Based on current research in progress, Babelscores wants to collaborate to solve problems through pragmatic and massif use of these techniques. The main vision BabelScores has is to create an interface between researchers, orchestras, creators, institutions and musicians that will allow these new usages to express their potentiality at their best.

2.4 INScore

INScore [26] is an environment for the design of augmented interactive music scores, opened to unconventional uses of music notation and representation, including real-time symbolic notation capabilities. It can be controlled in real-time using Open Sound Control [OSC] messages as well as using an OSC based scripting language, that allows designing scores in a modular and incremental way. INScore supports extended music scores, combining symbolic notation with arbitrary graphic objects. All the elements of a score (including purely graphical elements) have a temporal dimension (date, duration and tempo) and can be manipulated both in the graphic and time space. They can be synchronized in a master/slave relationship i.e. any object can be placed in the time space of another object, which may be viewed as "time synchronisation in the graphic space". As a result, a large number of operations can be performed in the time domain and in particular, moving a cursor on a score is simply achieved using the synchronization

mechanism and by moving this cursor in the time space. Time in INScore is both event-driven and continuous [19], which makes it possible to design interactive and dynamic scores. The system is widely open to network uses [17]: it allows to use both local and remote resources (via HTTP), it provides a forwarding mechanism that allows scores to be distributed in real time over a local network.

INScore has built-in solutions for monitoring the position and the speed of cursors in an efficient way. Since common practice notation is rarely proportional (i.e. a 4/4 is not necessarily twice longer on the page than a 2/4 bar), a cursor moving at constant speed across a whole system will never accurately fall at each bar accurately according to musical time. To tackle this issue, INScore provides a mechanism to describe the relationship between the graphic and the time space in the form of "mappings" which are actually a relation (in the mathematical sense) between a set of graphic segments and a set of temporal segments. A graphic segment is described with two intervals: the first one on the x-axis, and the second one on the y-axis. With images, these intervals are expressed in pixel positions. A time segment is an interval expressed by rationals that represent dates in musical time (where 1 is a whole note). INScore mappings describe arbitrary relationships between time and any graphical resource. It is mainly used to solve the problem of non-proportionality of symbolic musical notation (see for example in Fig. 4 and Fig. 5).

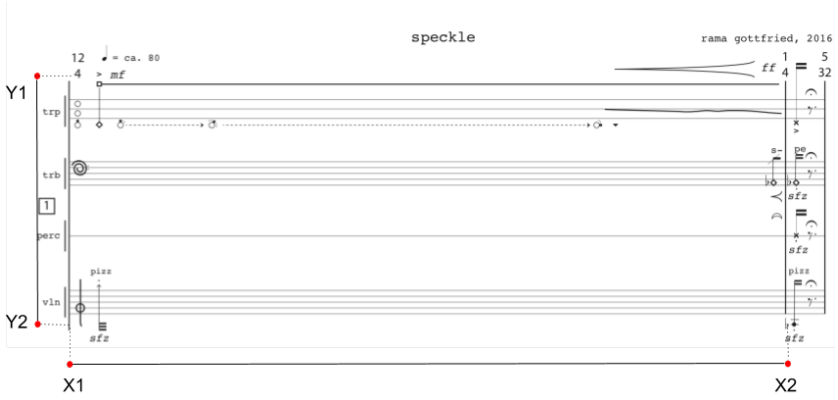


Fig. 4. One bar's coordinates expressed in pixel coordinates

These two segments (X1 - X2 and Y1 - Y2, see Fig. 4) define a rectangle corresponding to one bar in the score. The pixel coordinates of this rectangle are then attached to the duration of the bar, expressed in regular time signatures, with the following syntax: ([*one graphical segment on x axis [[one graphical segment on y axis[) ([one temporal segment[)*.

```
([144, 623[ [227, 745[) ([0/4, 1/4[)
# x1   x2     y1     y2      t1     t2
([635, 1203[ [705, 744[) ([1/4, 2/4[)
```

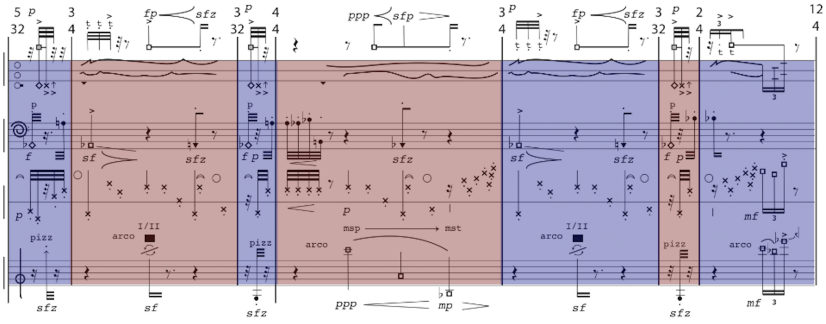


Fig. 5. Rectangles monitoring the position of the cursor: Time and space relationship is linear inside a box.

```
# x2+1    x3      y1      y2          t2      t3
```

The same procedure is then repeated with each bar (see Fig. 5). A specific INScore script allows you to draw mappings directly on the page and to retrieve them in a "nearly ready to use" form, with default time intervals that need to be specified.[15] This type of cursor seems to facilitate video part extraction by simply cropping the section of the screen corresponding to one particular instrument Figols - Fullscore (see corresponding separate part here).

3 Cursor Implementation

A cursor emphasizes the continuous unfolding of time while common practice musical notation implicitly relies on the rhythmic accents defined by bars and beats. The aim here is to find with INScore an efficient solution to incorporate animated cursors to a large number of pieces in the most automated and efficient possible way. Indeed, the *BabelBox* project consists in generalising to use of distributed notation and make it available to composers using those standard engraving softwares. From a given pdf score of the *BabelScores* corpus to its rendering on phone and tablet, one of the main issues will concern cursor implementation. Recent strategies for score distribution (such as pieces realised with the Decibel Score Player or the quintet.net/drawsocket server) tend let a score scroll over a fixed cursor. Whilst this solution remains envisageable, the BabelBox realisation achieved so far tend to use a moving cursor over fixed images, with page turns.

3.1 Cursors without INScore : Standard Engraving softwares and Bach

If all the pieces performed with SmartVox were composed in *Bach*, this environment, as its major antecedent *Open Music*, remains grounded in computer-aided composition

[15] The corresponding tools are available at the following address : https://github.com/grame-cncm/inscore/tree/dev/scripts/Tools/drawmap

rather than score engraving. Objects like *bach.roll* or *bach.score* are well-suited to animation and provide built-in cursor support, but they cannot compete (graphically) with the engraving capacities of dedicated softwares like Finale or Sibelius.

Latest versions of Sibelius (7.5 and 8.0 offer the possibility of exporting a video of the score with a cursor following the beats with the metronomic parameters marked in the score. This possibility could facilitate the creation of the video material needed for the *BabelBox*. However, the procedure demands a well configured MIDI encoding of the tempo - including tempo changes - in the original file in order not to miss the syncronisation between the notated score and the sound recording. Finale and Sibelius offer advanced MIDI controls that will have to be taken into account necessarily if this procedure is taken.

3.2 Continuous time with cursors, metrical time with blinkers

A cursor was also used in extracts of Emily Howard's Opera *To see the invisible*[16]. The main limitation of this type of representation resides in its impossibility to mark the beats implicitly expressed by time signatures. According to Richard Baker - who conducted the opera with the help of *SmartVox* in the second scene -, dynamic representations conveying the accent meant by the arm of a conductor might be preferable to the more linear trajectory of a cursor (see the "bouncing ball" for demonstration).

This "Bouncing ball" type of representation was introduced at Tenor 2018 by Pedro Louzeiro [23] and Slavko Zagorac [22]. It presents this advantage over cursors mark upbeats and downbeats, like the arm of a conductor, but one may argue that this form of representation might be disturbing for the eye,

Another strategy for marking the bar's beats more clearly might be to make a static cursor blink on each beat, with different colour for upbeat and downbeat (see examples realised on compositions by Brian Ferneyhough and Daniel Figols).

4 Conclusion

This paper presents the early stages of a research project which could help musicians worldwide[17] access distributed/animated notation easily. Indeed for pieces for chamber groups with electronics or video for instance, synchronisation through the browser of the performer's phone/tablet/computer seem a costless and promising way of making music. Solutions such as the BabelBox however, - local NPMs - only constitute a temporary solution if we acknowledge the exponential growth of the internet. In a few years' time, simple urls such as www.smartvox.eu should be sufficiently robust to let performers access the same piece via different terminals and through the internet directly, even in concert situations.[18]

[16] https://snapemaltings.co.uk/concerts-history/aldeburgh-festival-2018/to-see-the-invisible/

[17] BabelScores now has partnerships with the world's most prestigious universities, see https://www.babelscores.com/partners

[18] Rather than a static address (such as http://37.59.101.205:8000/), a bootstrap or a web-application may be more appropriate here.

INScore allows for the precise temporal control of animated cursors in the graphical domain. Thanks to its OSC support, it is possible to control the cursor's position via automations in Ableton Live.[19] This setup suits particularly well sound-to-visual synchronisation as in Julien Malaussena's piece, in which the cursor follows the audio recording. This methods however needs to be compared with the tools exposed in Chapter 3.1, which might allow for faster results.

Questions must be addressed regarding what is easier to do in INScore directly, in INScore with automation in *Ableton*, or in a video editor (Da Vinci resolve, Adobe Premiere, Final Cut...), according to what is sought:

- cursor with time given by the score : INScore (rectangles)
- cursor with time given by a recording of the piece : INScore automated in Ableton Max-for-Live?
- part extraction, two system per page layout : DaVinci ?

Once these practical problems elucidated, a more definitive solution for score elaboration workflow will be defined. Fig. 6 presents its current state.

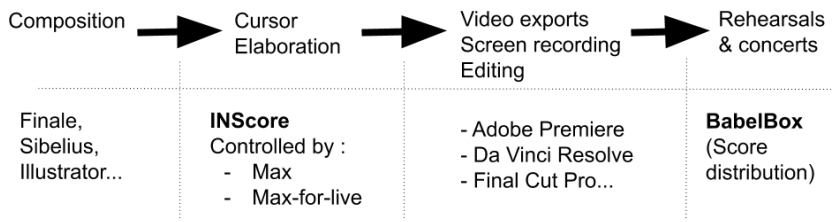


Fig. 6. Score elaboration workflow.

References

1. J. Bell, "Audio-scores, a resource for composition and computer-aided performance," Ph.D. dissertation, Guildhall School of Music and Drama, 2016. [Online]. Available: `http://openaccess.city.ac.uk/17285/`
2. J. Bell and B. Matuszewski, "SmartVox. A web-based distributed media player as notation tool for choral practices," in *Proceedings of the 3rd International Conference on Technologies for Music Notation and Representation (TENOR)*. Coruña, Spain: Universidade da Coruña, 2017.
3. J. Bell, "AUDIOVISUAL SCORES AND PARTS SYNCHRONIZED OVER THE WEB," in *TENOR 2018*, Montreal, France, 2018. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-01779806`
4. S. Zagorac and P. Alessandrini, "ZScore: A Distributed System For Integrated Mixed Music Composition and Performance," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 62–70.

[19] Via Max-for-Live, see https://youtu.be/rLy8DW_p2JE for demonstration.

5. C. Hope, A. Wyatt, and D. Thorpe, "Scoring an Animated Notation Opera – The Decibel Score Player and the Role of the Digital Copyist in 'Speechless'," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 193–200.
6. C. Hope, L. Vickery, A. Wyatt, and S. James, "The DECIBEL Scoreplayer - A Digital Tool for Reading Graphic Notation," in *Proceedings of the First International Conference on Technologies for Music Notation and Representation – TENOR'15*, M. Battier, J. Bresson, P. Couprie, C. Davy-Rigaux, D. Fober, Y. Geslin, H. Genevois, F. Picard, and A. Tacaille, Eds., Paris, France, 2015, pp. 58–69.
7. G. Hajdu and N. Didkovsky, "MaxScore: Recent Developments," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 138–146.
8. P. Louzeiro, "Improving Sight-Reading Skills through Dynamic Notation – the Case of Comprovisador," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 55–61.
9. A. Agostini and D. Ghisi, "BACH: an environment for computer-aided composition in Max," in *Proceedings of the 38th International Computer Music Conference (ICMC)*, Ljubljana, Slovenia, 2012.
10. C. Hope, "Electronic Scores for Music: The Possibilities of Animated Notation," *Computer Music Journal*, vol. 41, no. 3, pp. 21–35, 2017.
11. L. Vickery, "Some Approaches to Representing Sound with Colour and Shape," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 165–173.
12. R. Gottfried and J. Bresson, "Symbolist: An Open Authoring Environment for User-Defined Symbolic Notation," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 111–118.
13. J. Bresson, C. Agon, and G. Assayag, "OpenMusic – Visual Programming Environment for Music Composition, Analysis and Research," in *ACM MultiMedia (MM'11)*, Scottsdale, United States, 2011. [Online]. Available: `https://hal.archives-ouvertes.fr/hal-01182394`
14. D. G. Nathan Magnus, "Musician Assistance and Score Distribution (MASD)," in *Proceedings of The International Conference on New Interfaces for Musical Expression – NIME'2012*. Ann Arbor.: University of Michigan, 2012.
15. J. Bean, "denm (dynamic environmental notation for music): Introducing a Performance-Centric Musical Interface" in *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR)*. Ann Paris, France, 2015.
16. C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
17. D. Fober, Y. Orlarey, and S. Letz. Towards dynamic and animated music notation using inscore. In V. Ciciliato, Y. Orlarey, and L. Pottier, editors, *Proceedings of the Linux Audio Conference — LAC 2017*, pages 43–51, Saint Etienne, 2017. CIEREC.
18. D. Fober, G. Gouilloux, Y. Orlarey, and S. Letz, "Distributing Music Scores to Mobile Platforms and to the Internet using INScore," in *Proceedings of the Sound and Music Computing conference — SMC'15*, 2015, pp. 229–233. [Online]. Available: `inscore-web-SMC15.pdf`
19. D. Fober, Y. Orlarey, and S. Letz. Inscore time model. In *Proceedings of the International Computer Music Conference*, pages 64–68, 2017.

20. L. Vickery, "Hybrid Real/Mimetic Sound Works," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'16*, R. Hoadley, C. Nash, and D. Fober, Eds. Cambridge, UK: Anglia Ruskin University, 2016, pp. 19–24.
21. N. Schnell and S. Robaszkiewicz, "Soundworks – A playground for artists and developers to create collaborative mobile web performances," in *Proceedings of the first Web Audio Conference (WAC)*, Paris, France, 2015.
22. S. Zagorac and P. Alessandrini, "ZScore: A Distributed System For Integrated Mixed Music Composition and Performance," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 62–70.
23. P. Louzeiro, "Improving Sight-Reading Skills through Dynamic Notation – the Case of Comprovisador," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'18*, S. Bhagwati and J. Bresson, Eds. Montreal, Canada: Concordia University, 2018, pp. 55–61.
24. J. McPherson, "Turning-Score Automation for Musicians" in *Honours report, University of Canterbury*, New Zeland, 1999.
25. T. C. Bell, A. Church, J. McPherson, D. Bainbridge , "Page turning and image size in a digital music stand" in *International Computer Music Conference*, Barcelona, Spain, 2005.
26. D. Fober, Y. Orlarey, and S. Letz, "INScore – An Environment for the Design of Live Music Scores," in *Proceedings of the Linux Audio Conference – LAC 2012*, 2012, pp. 47–54. [Online]. Available: `INScore-ID12-2.pdf`

Methods and Datasets for DJ-Mix Reverse Engineering

Diemo Schwarz[1] and Dominique Fourer[2] *

[1] Ircam Lab, CNRS, Sorbonne Université, Ministère de la Culture, Paris, France
[2] IBISC, Université d'Évry-Val-d'Essonne/Paris-Saclay, Évry, France
schwarz@ircam.fr

Abstract. DJ techniques are an important part of popular music culture. However, they are also not sufficiently investigated by researchers due to the lack of annotated datasets of DJ mixes. Thus, this paper aims at filling this gap by introducing novel methods to automatically deconstruct and annotate recorded mixes for which the constituent tracks are known. A rough alignment first estimates where in the mix each track starts, and which time-stretching factor was applied. Second, a sample-precise alignment is applied to determine the exact offset of each track in the mix. Third, we propose a new method to estimate the cue points and the fade curves which operates in the time-frequency domain to increase its robustness to interference with other tracks. The proposed methods are finally evaluated on our new publicly available DJ-mix dataset. This dataset contains automatically generated beat-synchronous mixes based on freely available music tracks, and the ground truth about the placement of tracks in a mix.

1 Introduction

Understanding DJ practices remains a challenging important part of popular music culture [2,4]. The outcomes from such an understanding are numerous for musicological research in popular music, cultural studies on DJ practices and critical reception, music technology for computer support of DJing, automation of DJ mixing for entertainment or commercial purposes, and others. In order to automatically annotate recorded mixes, several components are required:

Identification of the contained tracks (e.g. fingerprinting) to obtain the playlist,
Alignment to determine where in the mix each track starts and stops,
Time-scaling to determine what speed changes were applied by the DJ to achieve beat-synchronicity,
Unmixing to estimate the cue regions where the cross-fades between tracks happen, the curves for volume, bass and treble, and the parameters of other effects (compression, echo, etc.),
Content and metadata analysis to derive the genre and social tags attached to the music to inform about the choices a DJ makes when creating a mix.

* The ABC_DJ project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 688122.

The first and last of these components have been addressed by recent MIR research. For alignment, time-scaling, and unmixing we propose a method based on multi-scale correlation, dynamic time warping, and time-frequency gain curve estimation to increase its robustness to interferences with other tracks. To come closer to actual DJ practices, we can retrieve the alignment and volume curves from example DJ mixes, and then combine them with content and genre information to investigate the content-dependent aspects of DJ mix methods.

As a working definition, we can roughly distinguish three levels of mixing:

Level 1, *broadcast mixing*, is a simple volume cross fade without paying attention to changing content (as performed by consumer audio players such as iTunes, or in a broadcast context).

Level 2, *lounge mixing*, is beat-synchronous mixing with adaptation of the speed of the tracks and possibly additional EQ fades, while playing the tracks mostly unchanged.

Level 3, *performative mixing*, is using the DJ deck as a performance instrument by creative use of effects, loops, and mashups with other tracks.

This paper addresses the level 1 and 2 cases, while level 3 can blur the identifiability of the source tracks.

2 Related Work

Rather than on DJ-mixing, existing work has focused on the field of *studio mixing*, where a stereo track is produced from multi-track recordings and software instruments by means of a mixing desk or DAW [4, 15, 16, 18]. It has produced ground truth databases [6] and crowd-sourced knowledge generation [7] with some overlap with DJ mixing. However, when seeing the latter as the mixing of only two source tracks, the studied parameters and influencing factors differ too much from what is needed for DJ mixing. There is quite some existing work on methods to help DJs to produce mixes [1,3,5,9,12,14,17], but much less regarding information retrieval from recorded mixes, with the exception of content-based analysis of playlist choices [13], track boundaries estimation in mixes [10,20], and the identification of the tracks within the mix by fingerprinting [24]. To this end, Sonnleitner et. al. provide an open dataset[3] of 10 dance music mixes with a total duration of 11 hours and 23 minutes made of 118 source tracks. The included playlists contain hand-annotated time points with relevant information for fingerprinting, namely the approximate instant when the next track is present in the mix. Unfortunately, this information is not accurate enough for estimating the start point of the track in the mix. As a result, it cannot be used for our aims of DJ mix analysis and let alone reverse engineering.

Barchiesi and Reiss [2] first used the term *mix reverse engineering* (in the context of multi-track studio mixing) for their method to invert linear processing (gains and delays, including short FIR filters typical for EQ) and some dynamic processing parameters (compression), of interest for our aim of DJ unmixing.

[3] http://www.cp.jku.at/datasets/fingerprinting

Ramona and Richard [19] tackle the unmixing problem for radio broadcast mixes, i.e. retrieving the fader positions of the mixing desk for several known input signals (music tracks, jingles, reports), and one unknown source (the host and guests' microphones in the broadcast studio). They model the fader curves as a sigmoid function and assume no time-varying filters, and no speed change of the sources, which is only correct in the context of radio broadcast. These two latter references both assume having sample-aligned source signals at their disposal, with no time-scaling applied, unlike our use-case where each source track only covers part of the mix, can appear only partially, and can be time-scaled for beat-matched mixing. There is rare work related to analysis [8] and inversion of non-linear processing applied to the signal such as dynamic-range compression [11] which remains challenging and full of interest for unmixing and source separation.

Hence, this work realizes our idea first presented in [21], by applying it to a large dataset of generated DJ mixes [22]. It already inspired work on a variant of our unmixing method based on convex optimization, and a hand-crafted database [26].

3 DJ Mix Reverse Engineering

The input of our method is the result of the previous stage of identification and retrieval on existing DJ mixes or specially contrived databases for the study of DJ practices. We assume a recorded DJ mix, a playlist (the list of tracks played in the correct order), and the audio files of the original tracks. Our method proceeds in five steps, from a rough alignment of the concatenated tracks with the mix by DTW (section 3.1), that is refined to close in to sample precision (section 3.2), then verified by subtracting the track out of the mix (section 3.3), to the estimation of gain curves (section 3.4) and cue regions (section 3.5).

3.1 Step 1: Rough Alignment

Rough alignment uses the Mel Frequency Cepstral Coefficients (MFCC) of the mix $X(k, c)$ (k being the mix frame index and $c \in \{1, 2, ..., 13\}$ the Mel frequency index) and the concatenated MFCCs of the I tracks $S(l, c) = (S_1 \dots S_I)$ as input (window size 0.05 s, hop size 0.0125 s), l being the frame index of the concatenated matrix S. The motivation for MFCC is that the representation should be compact, capture perceptually important attributes of the signals, and be robust against possible pitch changes from time-scaling of the source tracks in the DJ mix. MFCCs achieve this with only 13 coefficients and modeling of the spectral envelope, which captures rhythm and timbre, but not pitch. This is not the case for a discrete Fourier-based representation with 10 or 100x as many coefficients, where the match of spectral peaks of the tonal components would be degraded. Since the tracks are almost unchanged in level 2 mixes, Dynamic Time Warping (DTW) [25] can latch on to large valleys of low distance, although the fade regions in the mix are dissimilar to either track, and occur separately

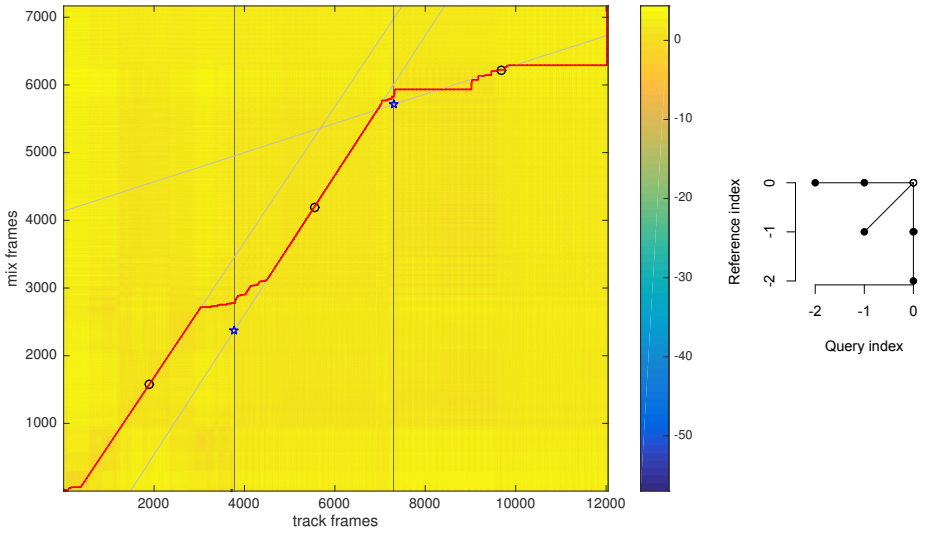


Fig. 1: Left: DTW distance matrix, alignment path (red), track boundaries (vertical lines), found slope lines anchored on track mid-point (circles), and estimated track start (blue marks) on an artificial DJ mix of 3 tracks from our dataset. Right: extended DTW neighbourhood.

in $S(l, c)$. To ease catching up with the shorter time of the mix, we provide a neighborhood allowing the estimated alignment path to perform larger vertical and horizontal jumps, shown in Fig. 1 (right).

The DTW alignment path not only gives us the relative positioning of the tracks in the mix, but also their possible speed change, applied by the DJ to achieve beat-synchronous mixing, see Fig. 1 (left): First, we estimate the speed factor, assuming that it is constant for each track, by calculating the mean slope of the alignment path in a window of half the track length, centred around the middle of the track. Then, the intersections of the slope lines with the track boundaries in $S(l, c)$ provide an estimate of the frame start of the tracks in the mix. The start position expresses the offset of the start of the full source track with respect to the mix, and not the point from where the track is present in the mix. Since the source tracks are mixed with non-zero volume only between the cue-in and cue-out regions, the track start point can be negative.

3.2 Step 2: Sample Alignment

Given the rough alignment and the speed estimation provided by DTW, we then search for the best sample alignment of the source tracks. To this end, we first time-scale the source track's signal according to the estimated speed factor using resampling. We then shift a window of the size of an MFCC frame, taken from the middle of the time-scaled track, around its predicted rough frame position in the mix. The best time shift is simply provided by the maximum of the cross-correlation computed between the mix and the track. Please note that this

process is not directly applied during the step 1 due to the high computational cost. The sample alignment considers a maximal delay equal to the size of a window and can be computed in a reasonable time.

3.3 Step 3: Track Removal

The success of the sample alignment can be verified by subtracting the aligned and time-scaled track signal from the mix for which a resulting drop in the root-mean-square (RMS) energy is expected. This method remains valid when the ground truth is unknown or inexact. Fig. 2 illustrates the result of track removal applied on a mix in our dataset. We can observe that the resulting instantaneous RMS energy of the mix (computed on the size of a sliding window) shows a drop of about 10 dB. A short increase is also observed during the fades where the suppression gradually takes effect.

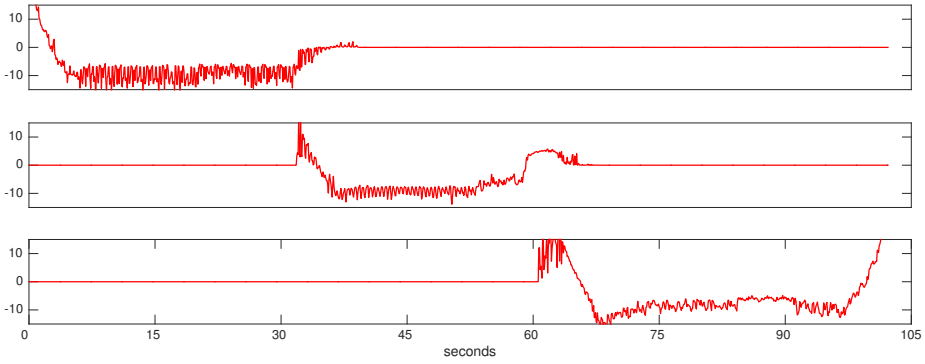


Fig. 2: Resulting RMS energy (in dB) after the subtraction of each track from a mix including fades. Each source track signal is padded with zeros to obtain the same duration than the mix.

3.4 Step 4: Volume Curve Estimation

We introduce a novel method based on the time-frequency representation of the signal to estimate the volume curve applied to each track to obtain the mix. Given the discrete-time mix signal denoted $x(n)$ and the constituent sample-aligned and time-scaled tracks $s_i(n)$, we aim at estimating the mixing function $a_i(n)$ as:

$$x(n) = \sum_{i=1}^{I} a_i(n) s_i(n) + b(n) \quad , \forall n \in \mathbb{Z} \tag{1}$$

where $b(n)$ corresponds to an additive noise signal.

From a "correctly" aligned track s_i, its corresponding volume curve $\hat{a}_i$ is estimated using the following steps:

1. we compute the short-time Fourier transforms (STFT) of x and s_i denoted $S_i(n,m)$ and $X(n,m)$ (n and m being respectively the time and frequency indices)
2. we estimate the volume curve at each instant n by computing the median of the mix/track ratio computed along the frequencies $m' \in \mathbb{M}$, where $\mathbb{M}$ is the set of frequency indices where $|S_i(n,m')|^2 > 0$, such as:

$$\hat{a}_i(n) = \begin{cases} \text{median}\left(\frac{|X(n,m')|}{|S_i(n,m')|}\right)_{\forall m' \in \mathbb{M}} & \text{if } \exists m' \text{ s. t. } |S_i(n,m')|^2 > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

3. we optionally post-process $\hat{a}_i(n)$ to obtain a smooth curve by removing outliers using a second median filter for which a kernel size equal to 20 provides good results in practice.

The resulting volume curve can then be used to estimate the cue points (the time instants when a fading effect begins or stops) at the next step. An illustration of the resulting process is presented in Fig. 3.

3.5 Step 5: Cue Point Estimation

In order to estimate the DJ cue points, we apply a linear regression of $\hat{a}_i$ at the time instants located at the beginning and at the end of the resulting volume curve (when $\hat{a}_i(n) < \Gamma$, Γ being a threshold defined arbitrarily as $\Gamma = 0.7 \max(\hat{a})$). Assuming that a linear fading effect was applied, the cue points can easily be deduced from the two affine equations resulting from the linear regression. The four estimated cue points correspond respectively to:

1. n_1, the time instant when the fade-in curve is equal to 0
2. n_2, the time instant when the fade-in curve is equal to $\max(\hat{a}_i)$
3. n_3, the time instant when the fade-out curve is equal to $\max(\hat{a}_i)$
4. n_4, the time instant when the fade-out curve is equal to 0.

In order to illustrate the efficiency of the entire method (steps 4 and 5), we present in Fig. 3 the results obtained on a real-world DJ-mix extracted from our proposed dataset.

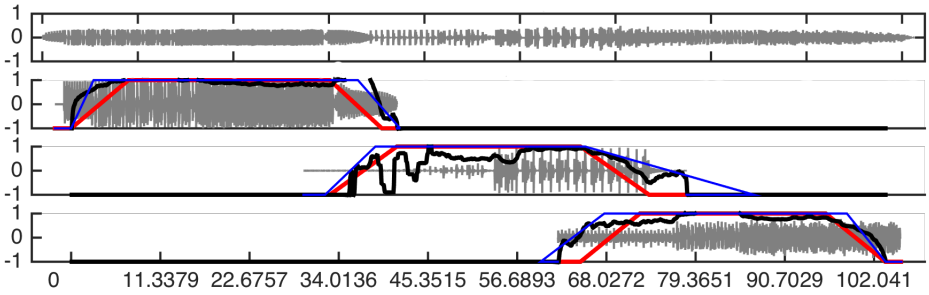


Fig. 3: Estimated volume curve (black), linear fades (blue), ground truth fades (red)

4 The *UnmixDB* Dataset

In order to evaluate the DJ mix analysis and reverse engineering methods described above, we created a dataset containing excerpts of open licensed dance tracks and their corresponding automatically generated mixes [22], available at https://zenodo.org/record/1422385. We use track excerpts of c.a. 40 seconds due to the high runtime and memory requirements, especially for the DTW that is of quadratic memory complexity.

Each mix is based on a playlist made of 3 track excerpts such that the middle track is embedded in a realistic context of beat-aligned linear cross fading to the other tracks. The first track's BPM is used as the seed tempo onto which the other tracks are adapted.

Each playlist of 3 tracks is mixed 12 times with combinations of 4 variants of effects and 3 variants of time scaling using the treatments of the *sox* open source command-line program.The 4 effects are:

none: no effect
bass: +6 dB bass boost using a low-shelving biquad filter below 100 Hz
compressor: heavy dynamics compression (ratio of 3:1 above -60 dB, -5 dB makeup gain)
distortion: heavy saturation with +20 dB gain

These effects were chosen to cover treatments likely to be applied to a DJ set (EQ, compression), and also to introduce non-linear treatments (distortion) to test the limits of re-engineering and unmixing methods.

The 3 timescale methods are:

none: no time scaling, ie. the tracks are only aligned on the first beat in the cue region and then drift apart
resample: linked time and pitch scaling by resampling (sox *speed* effect)
stretch: time stretching while keeping the pitch (sox *tempo* effect using WSOLA)

These 3 variants allow to test simple alignment methods not taking into account time scaling, and allow to evaluate the influence of different algorithms and implementations of time scaling.

The *UnmixDB* dataset contains the complete ground truth for the source tracks and mixes. For each mix, the start, end, and cue points of the constituent tracks are given with their BPM and speed factors. Additionally, the song excerpts are accompanied by their cue region and tempo information.

Table 1 shows the size and basic statistics of one of the six parts of the dataset (the one used for our evaluation). We also publish the Python source code to generate the mixes, such that other researchers can create test data from other track collections or other variants.

Our DJ mix dataset is based on the curatorial work of Sonnleitner et. al. [24], who collected Creative-Commons licensed source tracks of 10 free dance music mixes from the *Mixotic* net label. We used their collected tracks to produce our track excerpts, but regenerated artificial mixes with perfectly accurate ground truth.

Number of tracks	37
Number of playlists	37, tracks per playlist 3, variants per playlist 12
Number of mixes	444
Duration of tracks [min]	avg 0.76, total 1016
Duration of mixes [min]	avg 1.78, total 2743
Tempo of tracks [bpm]	min 67, median 128, mean 140

Table 1: Basic statistics of the *UnmixDB* dataset.

5 Evaluation

We applied the DJ mix reverse engineering method on our *UnmixDB* collection of mixes and compared the results to the ground truth annotations. To evaluate the success of our method we defined the following error metrics:

frame error: absolute error in seconds between the frame start time found by the DTW rough alignment (step 1, section 3.1) and the ground truth (virtual) track start time relative to the mix

sample error: absolute error in seconds between the track start time found by the sample alignment (step 2, section 3.2) and the ground truth track start time relative to the mix

speed ratio: ratio between the speed estimated by DTW alignment (step 1, section 3.1) and the ground truth speed factor (ideal value is 1)

suppression ratio: ratio of time where more than 15 dB of signal energy could be removed by subtracting the aligned track from the mix, relative to the time where the track is fully present in the mix, i.e. between fade-in end and fade-out start (step 3, section 3.3, bigger is better)

fade error: the total difference between the estimated fade curves (steps 4 and 5, sections 3.4 and 3.5) and the ground truth fades. This can be seen as the surface between the 2 linear curves over their maximum time extent. The value has been expressed in dB s, i.e. for one second of maximal difference (one curve full on, the other curve silent), the difference would be 96 dB.

Figures 4–10 show the quartile statistics of these metrics, broken down by the 12 mix variants (all combinations of the 3 time-scaling methods and 4 mix effects). The sample alignment results given in Fig. 6 and Table 2 show that the ground truth labels can be retrieved with high accuracy: the median error is 25 milliseconds, except for the mixes with distortion applied, where it is around 100 ms. These errors can already be traced back to the rough alignment (section 3.1): Fig. 4 shows that it is not robust to heavy non-linear distortion, presumably because the spectral shape changes too much to be matchable via MFCC distances. This error percolates to the speed estimation (Fig. 5), and sample alignment.

The track removal time results in Fig. 8 show sensitivity to the bass and distortion effect (because both of these introduce a strong additional signal component in the mix that is left as a residual when subtracting a track), and also perform less well for time-scaled mixes.

The fade curve volume error in Fig. 10 shows a median of around 5 dB s, which corresponds to a very good average dB distance of 0.3 dB, considering that the fades typically last for 16 seconds.

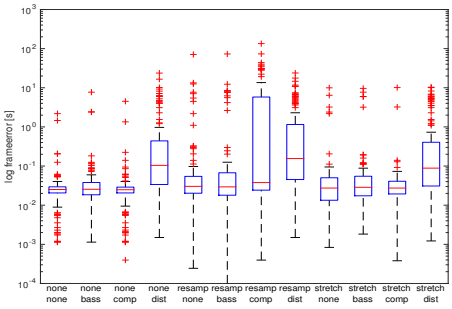


Fig. 4: Box plot of absolute error in track start time found by DTW per variant.

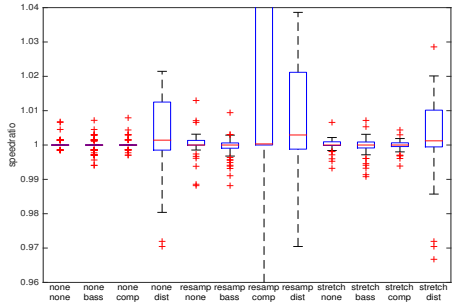


Fig. 5: Box plot of ratio between estimated and ground truth speed per variant.

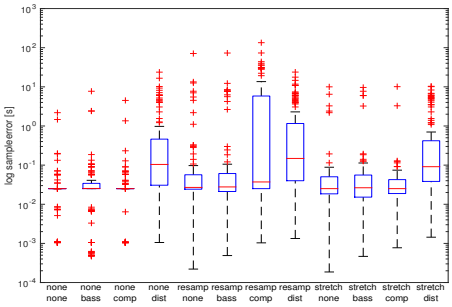


Fig. 6: Box plot of absolute error in track start time found by sample alignment per variant.

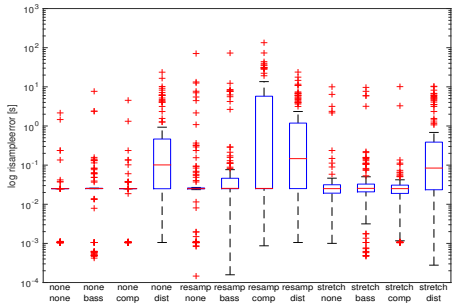


Fig. 7: Box plot of absolute error in track start time found by sample alignment when re-injecting ground truth speed.

While developing our method, we noticed the high sensitivity of the sample alignment and subsequent track removal (steps 2 and 3, sections 3.2 and 3.3) on the accuracy of the speed estimation. This is due to the resampling of the source track to match the track in the mix prior to track removal. Even an estimation error of a tenth of a percent results in desynchronisation after some time.

To judge the influence of this accuracy, we produced a second set of the *sample error* and *suppression ratio* metrics based on a run of steps 2 and 3 with the ground truth speed re-injected into the processing. The rationale is that the speed estimation method could be improved in future work, if the resulting reductions of error metrics are worthwhile. Also note that the tempo estimation is inherently inaccurate due to it being based on DTW's discretization into MFCC frames. In mixes with full tracks, the slope can be estimated more accurately than with our track excerpts simply because more frames are available.

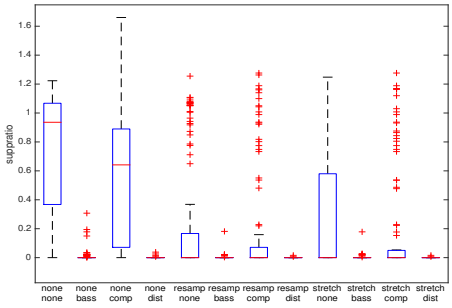


Fig. 8: Box plot of ratio of removal time (bigger is better) per variant.

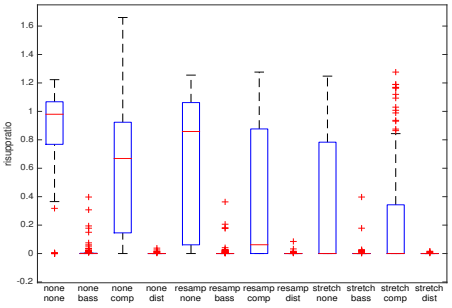


Fig. 9: Box plot of ratio of removal time when re-injecting ground truth speed (bigger is better) per variant.

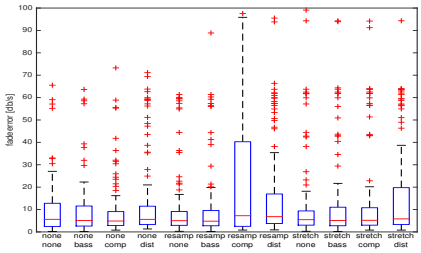


Fig. 10: Box plot of volume difference of fades per variant.

Figures 7 and 9 show the quartile statistics of the sample error and suppression ratio with re-injected ground truth speed. We can see how most variants are improved in error spread for the former, and 4 variants are greatly improved for the latter, confirming the sensitivity of the track removal step 3 on the speed estimation.

	mean	std	min	median	max
none none	0.0604	0.2469	0.0010	0.0251	2.1876
none bass	0.1431	0.7929	0.0005	0.0254	7.7191
none compressor	0.0806	0.4424	0.0010	0.0251	4.4995
none distortion	1.3376	3.3627	0.0011	0.1042	23.7610
resample none	1.1671	7.0025	0.0002	0.0270	71.0080
resample bass	1.3337	7.2079	0.0005	0.0277	73.1192
resample compressor	6.8024	17.0154	0.0010	0.0372	134.2811
resample distortion	1.8371	3.8551	0.0013	0.1483	23.8355
stretch none	0.2502	1.1926	0.0002	0.0251	10.0048
stretch bass	0.3300	1.4249	0.0005	0.0264	9.6626
stretch compressor	0.1520	1.0025	0.0008	0.0251	10.1076
stretch distortion	1.0629	2.2129	0.0014	0.0911	10.3353
all	1.2131	6.2028	0.0002	0.0282	134.2811

Table 2: Statistics of absolute error in track start time found by sample alignment.

6 Conclusions and Future Work

The presented work is a first step towards providing the missing link in a chain of methods that allow the retrieval of rich data from existing DJ mixes and their source tracks. An important result is the validation using track removal in section 3.3 to compute a new metric for the accuracy of sample alignment. This metric can also be computed even without ground truth. A massive amount of training data extracted from the vast number of collections of existing mixes could thus be made amenable to research in DJ practices, cultural studies, and automatic mixing methods. With some refinements, our method could become robust and precise enough to allow the inversion of fading, EQ and other processing [2, 19]. First, the obtained tempo slope could be refined by searching for sample alignment at several points in one source track. This would also extend the applicability of our method to mixes with non-constant tempo curves. Second, a sub-sample search for the best alignment should achieve the neutralisation of phase shifts incurred in the mix production chain. Various improvements of DTW alignment are possible: relaxed endpoint conditions [23] could allow pre-alignment per track and thus reduce the memory requirements, and better handle the partial presence of tracks in the mix. Furthermore, the close link between alignment, time-scaling, and unmixing hints at the possibility of a joint and possibly iterative estimation algorithm, maximising the match in the three search spaces simultaneously. In further experiments, we will test the influence of other signal representations (MFCC, spectrum, chroma, scattering transform) on the results, and could extend the *UnmixDB* dataset by other effects commonly used in DJing (cuts and boost on bass, mid and high EQ).

References

1. Felipe Aspillaga, J. Cobb, and C-H Chuan. Mixme: A recommendation system for DJs. In *ISMIR*, October 2011.
2. Daniele Barchiesi and Joshua Reiss. Reverse engineering of a mix. *Journal of the Audio Engineering Society*, 58(7/8):563–576, 2010.
3. Rachel M. Bittner, Minwei Gu, Gandalf Hernandez, Eric J. Humphrey, Tristan Jehan, Hunter McCurry, and Nicola Montecchio. Automatic playlist sequencing and transitions. In *ISMIR*, October 2017.
4. Brett Brecht De Man, R; King, and J. D. Reiss. An analysis and evaluation of audio features for multitrack music mixtures. In *ISMIR*, 2014.
5. Dave Cliff. Hang the DJ: Automatic sequencing and seamless mixing of dance-music tracks. Technical report, Hewlett-Packard Laboratories, 2000. HPL 104.
6. Brecht De Man, Mariano Mora-Mcginity, György Fazekas, and Joshua D Reiss. The open multitrack testbed. In *Audio Engineering Society Convention 137*, 2014.
7. Brecht De Man and Joshua D Reiss. Crowd-sourced learning of music production practices through large-scale perceptual evaluation of mixes. *Innovation in Music II*, page 144, 2016.
8. Dominique Fourer and Geoffroy Peeters. Objective characterization of audio signal quality: Application to music collection description. In *Proc. IEEE ICASSP*, pages 711–715, March 2017.

9. Tsuyoshi Fujio and Hisao Shiizuka. A system of mixing songs for automatic DJ performance using genetic programming. In *6th Asian Design International Conference*, October 2003.
10. Nikolay Glazyrin. Towards automatic content-based separation of DJ mixes into single tracks. In *ISMIR*, pages 149–154, October 2014.
11. Stanislaw Gorlow and Joshua D Reiss. Model-based inversion of dynamic range compression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1434–1444, 2013.
12. Hiromi Ishizaki, Keiichiro Hoashi, and Yasuhiro Takishima. Full-automatic DJ mixing system with optimal tempo adjustment based on measurement function of user discomfort. In *ISMIR*, pages 135–140, 2009.
13. Thor Kell and George Tzanetakis. Empirical analysis of track selection and ordering in electronic dance music using audio feature extraction. In *ISMIR*, pages 505–510, 2013.
14. Adrian Kim, Soram Park, Jangyeon Park, Jung-Woo Ha, Taegyun Kwon, and Juhan Nam. Automatic DJ mix generation using highlight detection. In *Proc. ISMIR, late-breaking demo paper*, October 2017.
15. Jacob A Maddams, Saoirse Finn, and Joshua D Reiss. An autonomous method for multi-track dynamic range compression. In *DAFx*, 2012.
16. Stuart Mansbridge, Saorise Finn, and Joshua D Reiss. An autonomous system for multitrack stereo pan positioning. In *Audio Engineering Society Convention*, 2012.
17. Pablo Molina, Martín Haro, and Sergi Jordá. Beatjockey: A new tool for enhancing DJ skills. In *NIME*, pages 288–291. Citeseer, 2011.
18. Enrique Perez-Gonzalez and Joshua Reiss. Automatic gain and fader control for live mixing. In *Proc. IEEE WASPAA*, pages 1–4, October 2009.
19. Mathieu Ramona and Gaël Richard. A simple and efficient fader estimator for broadcast radio unmixing. In *Proc. Digital Audio Effects (DAFx)*, pages 265–268, September 2011.
20. Tim Scarfe, W Koolen, and Yuri Kalnishkan. Segmentation of electronic dance music. *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications*, 22(3):4, 2014.
21. Diemo Schwarz and Dominique Fourer. Towards Extraction of Ground Truth Data from DJ Mixes. In *Late-break Session of ISMIR*, Suzhou, China, October 2017.
22. Diemo Schwarz and Dominique Fourer. Unmixdb: A dataset for DJ-mix information retrieval. In *Late-break Session of ISMIR*, Paris, France, September 2018.
23. Diego Furtado Silva, Gustavo Enrique de Almeida Prado Alves Batista, Eamonn Keogh, et al. On the effect of endpoints on dynamic time warping. In *SIGKDD Workshop on Mining and Learning from Time Series, II*. Association for Computing Machinery-ACM, 2016.
24. Reinhard Sonnleitner, Andreas Arzt, and Gerhard Widmer. Landmark-based audio fingerprinting for DJ mix monitoring. In *ISMIR*, New York, NY, 2016.
25. Robert J Turetsky and Daniel PW Ellis. Ground-truth transcriptions of real music from force-aligned midi syntheses. In *ISMIR*, October 2003.
26. Lorin Werthen-Brabants. Ground truth extraction & transition analysis of DJ mixes. Master's thesis, Ghent University, Belgium, 2018.

Identifying Listener-informed Features for Modeling Time-varying Emotion Perception

Simin Yang[1], Elaine Chew[2], Mathieu Barthet[1]

[1] Centre for Digital Music, Queen Mary University of London, UK
{simin.yang, m.barthet}@qmul.ac.uk
[2] CNRS-UMR9912/STMS IRCAM, Paris, France
elaine.chew@ircam.fr

Abstract. Music emotion perception can be highly subjective and varies over time, making it challenging to find salient explanatory acoustic features for listeners. In this paper, we dig deeper into the reasons listeners produce different emotion annotations in a complex classical music piece in order to gain a deeper understanding of the factors that influence emotion perception in music performance. An initial study collected time-varying emotion ratings (valence and arousal) from listeners of a live performance of a classical trio; a follow-up study interrogates the reasons behind listeners' emotion ratings through the re-evaluation of several preselected music segments of various agreement levels informed from the initial study. Thematic analysis of the time-stamped comments revealed themes pertaining primarily to musical features of loudness, tempo, and pitch contour as the main factors influencing emotion perception. The analysis uncovered features such as instrument interaction, repetition, and expression embellishments, which are less mentioned in computational music emotion recognition studies. Our findings lead to proposals for ways to incorporate these features into existing models of emotion perception and music information retrieval researches. Better models for music emotion provide important information for music recommendation systems and applications in music and music-supported therapy.

Keywords: music and emotion, live performance, human computer interaction, thematic analysis

1 Introduction and Background

Music perception studies show that the same music can communicate a range of emotions that vary over time and across listeners [16, 27]. Time-continuous annotation of music enables to capture detailed localised emotion cues, and inter-rater differences can be studied by involving multiple annotators. Previous music emotion studies have evidenced correlations between musical attributes such as dynamics, tempo, mode, timbre, harmony, articulation, timbre, and emotion judgements [11, 15, 18]. In the Music Emotion Recognition (MER) field, several approaches have been proposed to map acoustic features to time-continuous emotional annotations [17, 26, 22]. Yet, little is known on the relative importance

of these features across listeners. Machine learning approaches for MER yielded improved performances overall through extensive testing of different feature sets (bag of audio words), however, these approaches are facing the issue of confounded model performances [1, 14]. In addition, most of the low-level acoustic features involved such as Mel-frequency cepstral coefficients (MFCCs) do not explain the underlying cognitive mechanisms [2, 5, 31].

The subjective nature of music emotion perception has also been less investigated [11, 30]. Traditional approaches to dynamic emotion recognition typically take the average of multi-rater annotations as "target" and discard inconsistent ratings; however, subjective ratings can make the average prone to reliability issues. The variability in rater agreement with the ground truth data may induce a natural upper bound for any algorithmic approach, thus a bottleneck of the MER system performance [13]; it might also lead to a systematic misrepresentation of emotion perception [10]. Such potential limits have also been discussed by in the context of the largest publicly available emotion dataset to date, (DEAM) [1], which provides multi-rater time-varying emotion annotations on over 1800 tracks. Since relatively low agreement between annotators has been found in this dataset, the authors propose as future perspective that *"instead of taking the average values of the emotional annotations as the ground truth and training a generalised model for predicting them, we might want to have a look at the raw annotations and investigate the difference across the annotators."*. This highlights the importance of inter-rater variability in MER researches. As emotion data acquisition can be really expensive and time-consuming, it would be a loss to ignore subjective information which may already exist in available emotion datasets.

In this paper, we present an empirical study aiming to better understand the factors that influence emotion judgements, by exploring time-varying music emotion ratings in a real classical music performance. After collecting emotional annotations from participants in a live context, we conducted exploratory research to find the most relevant features. This was done by asking participants to re-evaluate time-stamped emotion ratings and explain their choice. This provides us with factors related to emotion ratings that have a cognitive meaning. Initial thematic analysis [8] of the time-stamped explanations revealed themes pertaining primarily to musical features of loudness, tempo, and pitch contour as the main factors influencing emotion perception. The analysis also uncovered features such as instrument interaction, repetition, and expression embellishments which are less employed in computational music emotion models. With the recent advances in music information retrieval e.g. in source separation and instrument recognition, listener-informed features can potentially be incorporated for future MER research.

2 Data Acquisition and Statistical Analyses

2.1 Stimulus: Babajanian Piano Trio

In a previous study [35], we collected time-based emotion annotations in a live music performance setting. We chose the piece *Piano Trio in F# minor by Arno Babajanian* which was performed by a professional pianist, cellist and violinist. This piece contains widely disparate characters; as a result, it might express various emotion to participants over time and enable us to capture more explanations from different listeners' perspectives; also this piece is rarely known to the public, thus avoiding familiarity bias. 15 participants provided ratings of valence (degree of pleasantness) and arousal (degree of excitation) [25] which were collected using our web-based and smartphone-friendly app Mood Rater based on a previous framework for audience participation in live music [12]. The audio recording of the concert and emotion data logged on the server-side were synchronised thanks to timestamps. Previous analyses showed varied levels of inter-rater agreement [29], from very low agreement to significant agreement. These results lead us to conduct a follow-up study, which is described in the present paper, in order to better understand the factors influencing listeners judgements of valence and arousal in response to music.

In the follow-up study, we used the video recording of the first two movements of the performance, resulting in stimuli of 17 minutes in length. According to the score provided by the performers, the first movement is marked *Largo-Allegro expressivo-Maestoso*, meaning it is largely in a slow tempo with a faster middle part; the second is marked *Andante*, meaning it is performed at a walking pace. The piece could be segmented into 25 segments based on rehearsal marks on the score[3], lasting from 38 to 72 seconds. Considering the duration of the study for participants, we selected seven excerpts (Segment 5, 7, 12, 13, 14, 17) within the recording for reflective feedback. These excerpts last from 38 to 67 seconds and last 6 minutes in total. Selection of these seven excerpts was based on the diversity of music attributes represented by the stimuli (e.g., instrumentation, loudness, tempo), the diversity of agreement levels of emotion ratings among listeners to cover both commonalities and divergences in music emotion perception. The ICCs of these seven selected experts range from ICC=-0.13, $p > 0.05$ to ICC=0.67, $p < 0.05$ in both arousal and valence.

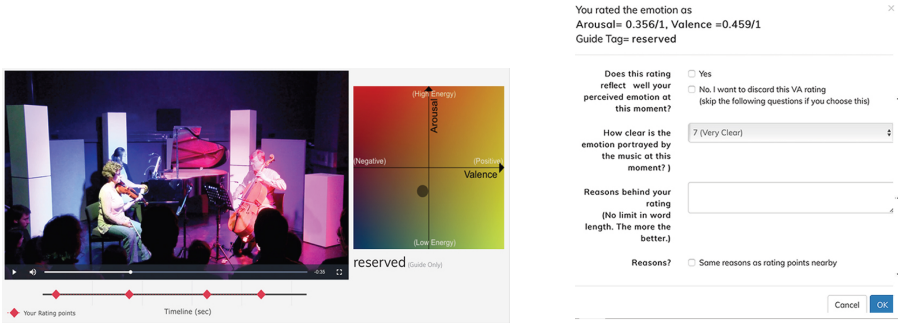
2.2 Procedure

The follow-up study consisted of a rating task followed by a reflective feedback task. Each participant was seated in front of a computer in a quiet sound-proofed room and interacted with a web-based application for stimulus delivery and data

[3] Rehearsal marks are used to identify specific points in a score to facilitate rehearsing. Many scores and parts have bar numbers, every five or ten bars, or at the beginning of each page or line. But as pieces and individual movements of works became longer (extending to several hundred bars), rehearsal marks became more practical in rehearsal, which provides a guideline to segment music.

acquisition. Sound stimuli were presented through headphones with the same sound level (Beyerdynamic DT 770 Pro). Participants were first introduced to the goal of the study, the Valence and Arousal (VA) space, and the self-report framework. Participants then followed a rating trial.

After the rating trial, participants rated the perceived emotion while watching the video recording. They could rate the perceived emotion whenever they perceived a change by clicking on the VA space presented next to the video. In particular, participants were informed that ratings were assumed constant until a change was made. Participants were allowed to pause or rewind the music as needed. For each click on the VA space, both the corresponding UTC timestamp and the corresponding time position of the video were recorded. In addition, corresponding emotion tags were shown below the VA space along upon clicking to help participants to use the VA space. These tags were selected based on [7, 36], which provide a set of normative emotional ratings for a large number of words in English. Participants were informed that these tags were only a guide and they could have their own interpretations of the VA space.



(a) Rating points review

(b) Pop-up window for rating confirmation

Fig. 1: Interfaces for reflective feedback task in the follow-up study (reflective condition)

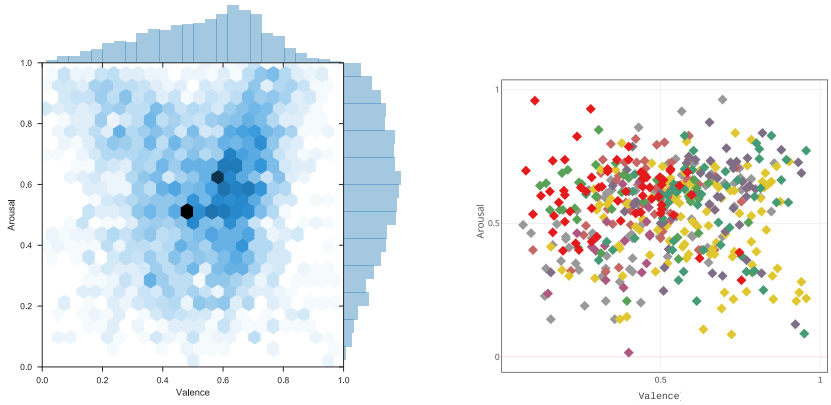
After the rating task, participants started the reflective feedback task. This task was designed for participants to review and confirm each emotion rating they had just given. As shown in Figure 1a, the emotion rating points (shown as red diamonds) were automatically displayed under the video on a synchronised timeline with the video time-slider. By hovering over the rating points, the corresponding VA ratings would be presented in the VA space on the right panel for reflective feedback. By clicking on each rating point, a pop-up window (Figure 1b) appeared for participants to confirm their rating and assess how clearly the emotion was perceived (from 1, very unclear, to 7, very clear). A comment box was provided to allow participants to provide reasons for their ratings using free descriptions. Participants were made aware that there were no right or wrong

answers and they were invited to report as much as possible. After the two tasks, participants completed a questionnaire to collect demographic information, as well as information such as music experience (Goldsmith Music Sophistication Index [24]). The duration of the whole study for each participant ranged from 1.5 to 2.5 hours.

2.3 Participants

21 participants (11 males and 10 females; age M= 28.8, SD=5.5; age range: 23-46 years) participated in the study. One participant stopped after reviewing the first 2 excerpts. Participants had varying degrees of music training (years of engagement in regular, daily practice of a musical instrument: >10 year: 11; 6-9 years: 1; 4-5 years: 1; 1-2 years: 3; 0 year: 5). All participants were current residents in the United Kingdom.

2.4 Explanatory Statistics of Collected Data



(a) Distribution of 3181 VA ratings over 17 minutes with histogram using hexagonal bins (Rating task)

(b) Distribution of 483 explanation ratings, rating points fell on separate segments are shown in different colours (Reflective feedback task)

Fig. 2: Distribution of collected data in the follow-up study (reflective condition)

Rating Task: Over the course of the live performance recording (17 minutes, 25 segments), 3181 VA emotion ratings were collected in total from the 21 participants (151± 96 per participant). Figure 2a depicts the distribution of all 3181 collected VA ratings. This figure shows that the collected data span all four quadrants of the VA space, which is in line with the varied expression within the piece. By comparing the time differences in UTC timestamp as well as those

of video recording timestamps, we found that 10 people rewound or paused the video during the rating process, and no one skipped or fast-forwarded the video. **Reflective feedback task:** 21 participants re-evaluated the 1098 VA ratings they have given on seven pre-selected segments. Among 1098 reviewed ratings, the participants gave explanations and clarity levels towards 471 ratings and categorised another 605 ratings as transition ratings, owing the same reasons than others. 8 participants discarded 23 previous ratings and 7 participants provided 12 new ratings. We collected 483 explanations (23±9 explanations per participant, 7000+ words in total) in total. From Figure 2b we can see that the ratings cover a fairly wide span of the VA space. Hence the explanations represent a broad coverage of emotional responses for the recorded live music performance.

2.5 Measuring Agreement in Participants Emotion Ratings

To quantify the agreement between participants, we computedthe Intra-class Correlation (ICC) [29] at rehearsal segment-level for participants' Valence and Arousal emotion ratings. Specifically, the case of two-way mixed, agreement, average-measures (ICC(2,k)) was adopted for estimating the reliability of the averaged ratings among listeners. Higher ICC values correspond to higher degrees of agreement among listeners, an ICC value of 1 indicates total agreement, while an ICC value of 0 represents random agreement. Negative ICC values are also possible, indicating systematic disagreement. As participants were informed that their emotion will be assumed unchanged until they sent a new rating, we resampled individual emotion ratings using a step function at 1Hz for the ICC calculation. The ICC results from both the initial study (live condition) [35] and the current study (reflective condition) are presented in Figure 3.

The ICC of both Arousal and Valence in reflective condition are higher than in the live condition. Possible reasons include: a higher focus and concentration for such an emotion rating task in the lab setting as a single participant compared to real-world live performance setting involving social interactions; the possibility to pause and rewind the videos; differences between groups of participants and larger sample size for ICC calculation in the present study.

3 Listener-informed Features for Music Emotions

3.1 Initial Thematic Analysis on Explanations towards Emotion Ratings

We examined participants' explanations using inductive (bottom-up) thematic analyses [8], a qualitative content analysis approach aiming to look closely into the text in order to find patterns of similar meaning, more than just using a simple count for frequencies of text occurrence.

483 time-stamped explanation data (comments) towards all seven music segments were imported into NVIVO 12 for analysis. Each of the explanation comment was first assigned one or multiple "codes" that identified a feature of the

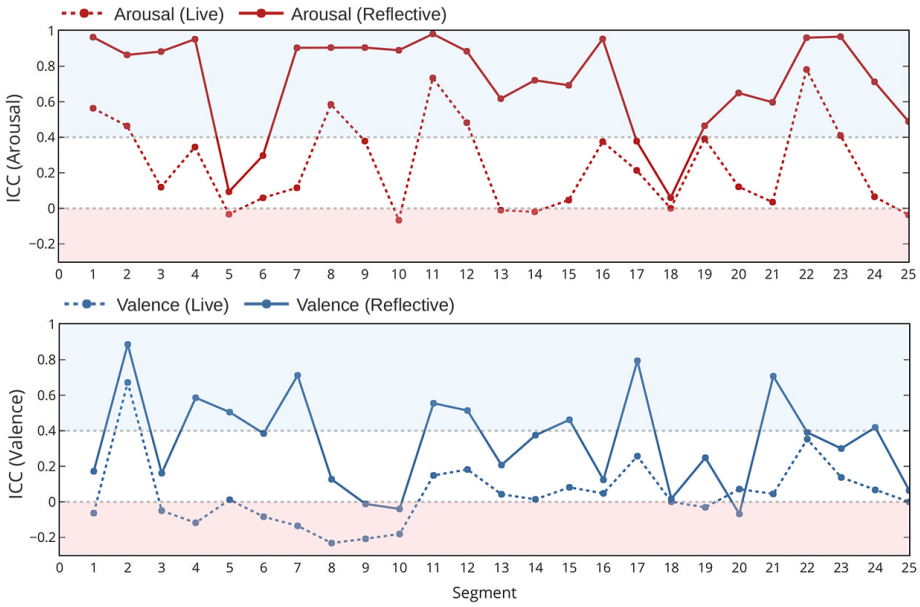


Fig. 3: Intra-class correlation (ICC) for Arousal (top) and Valence (bottom) in both the Live (dotted lines) and Reflective (plain lines) conditions

comment. Broader themes, which were not predetermined, were then obtained by refocusing the analysis at a broader perspective and collating all the relevant coded data within the identified themes.

Figure 4 presents the main themes and the associated codes with their number of occurrence. The occurrence of each code, counted in terms of the number of comments which referred to it, are attached next to each code. As we can see from Figure 4, ten key themes were obtained: **Dynamics, Rhythm, Melody, Harmony, Timbre, Instrument, Structure, Expression, Visuals cues**. It should be noted that some of the themes which emerged overlap as explanations are often multifaceted, such as between **Dynamics** and **Instrument**. In the following discussion, the following notation is used: N refers to the total number of codes for a (sub)theme, and C refers to the number of comments in which a code is found.

As shown in Figure 4, **Dynamics (N=209)** is the most frequently mentioned theme. In this piece, *loudness* (N=169) seems to have been the most salient feature behind participants' music emotion perception. References to **Rhythm (N=114), Harmony (N=114), Melody (N=113)** are also frequently made. Under these three themes, *tempo* (N=74), *pitch contour* (N=67), *mode(major, minor)* (N=50) emerged as three salient factors for music emotion perception. These themes are in line with previous music emotion perception studies which have shown the importance of dynamics, tempo, mode in music emotion perception. In addition, the following themes were found: **Instrument (N=177),**

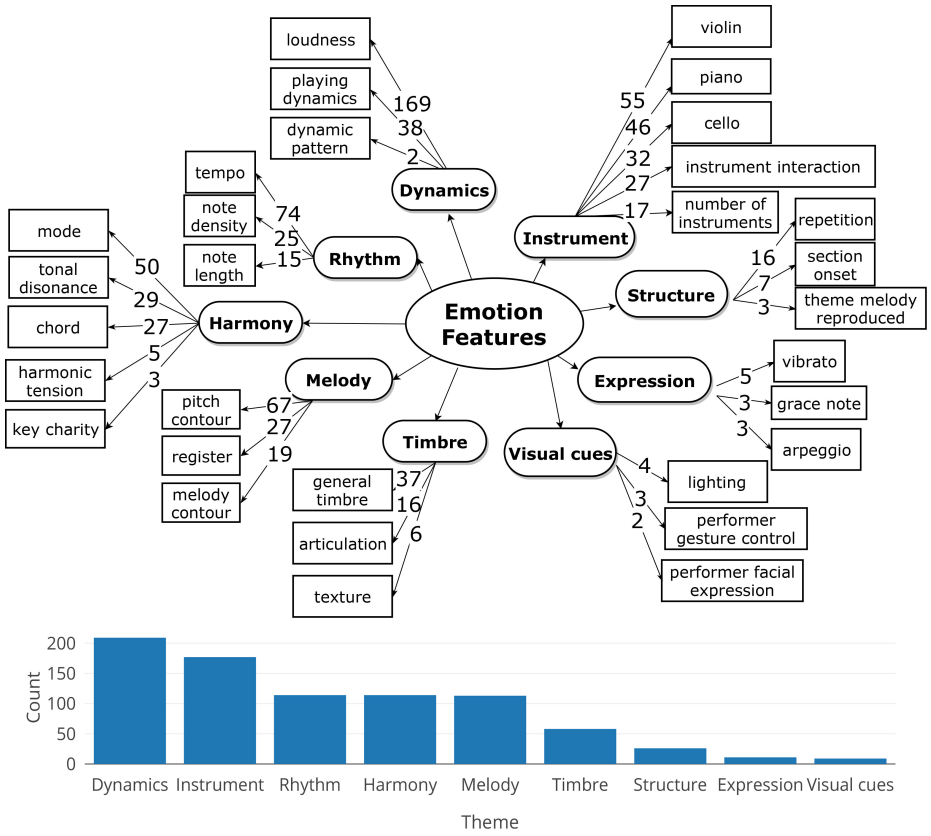


Fig. 4: Thematic analysis of audiences explanation comments

Structure (N=26), Expression (N=11), Visuals cues (N=9). Since these factors are less mentioned in computational emotion research, we discuss them into more details in the following.

Instrument (N=177) Under this theme, many people associated their emotion judgements with one specific instrument or multiple instruments. Violin (N=55), piano (N=46) and cello (N=32) were all frequently referred to for participants' emotional judgements. It provides an indication that some people pay attention to different instruments, which influence their perception of emotion. There are many parameters that performers can control and shape depending on the instrument, from loudness, tempo, timing, articulation to complex continuous aspects such as intonation, instrument timbrel control, and ornaments. Although similar levels of loudness can be reached with different instruments, they can be discriminated by their timbre, and timbre variations have been shown to be an important factor of expressiveness [4]. It can be assumed that performers' timbrel variations also influence the perception of emotion. Other than this, we also extracted codes relating to *instrument interaction* (N=27)

when participants referred to the music with a specific collaboration between multiple performers with multiple instruments, which is sensible as for much of the time music is played in ensembles for instance, the following cases that were mentioned by participants: 1. Multiple instruments are playing the same music melody, which affects the perception of arousal and valence (C=4) 2. The appearance of an instrument can lead to changes of emotion perception, e.g. *"First the violin and cello melty start more warm, then, the piano starts playing and energy increase.*" (C=5) 3. Two instruments were responding to each other e.g. *"strong notes alternated between piano and violin"* (C=3).

Structure (N=26) This theme refers to participants' comments on emotion referring to music structure. Supporting codes are *repetition, section onset, theme melody reproduced*. Participants associated their emotion with *repetition* (N=16) of specific music patterns, e.g. *"repetition of same melody accompanying increasing loudness and pitch build up the emotion"*. Transition points, or onsets of a new section within the music, are also possibly lead to the emotion change(N=7). Participants also associated emotion change with the reappearance of theme melody at a given point within the performance (N=3).

Expression (N=11) We categorised supporting codes that were referring to specific music embellishments under this theme. Music embellishments can be obtained by adding notes or producing particular variations to decorate the main music line (or harmony). In particular, people associated the emotion changes with vibratos in violin (N=5), grace notes in piano (N=3) and arpeggios in piano (N=3). Interestingly, these specific factors are mentioned by people with over ten years' music training in violin and piano respectively, and it indicates that people might pay more attention to the instrument they have expertise in playing for emotion perception.

Visual cues (N=9) As participants rated video recordings, some actively reported reasons from the visual perspective, even if this was not mentioned in the task. Participants mentioned the lighting influenced their emotion perceptions. In particular, participants associated the decrease of arousal as the lights turned dark in the final examined segment (N=4). People also referred to the motions of performer gesture, such as bow movement on cello and violin, as reasons for emotion judgements (N=5). Besides, participants mentioned the facial expressions they observed from the performers as reasons, e.g. *"cellist's face looks very expressive, face screws up"*.

3.2 Insights for Building MER Models and MIR

The identification of appropriate and well-functioning features is one of the most important targets in Music information retrival (MIR) researches. Based on our current findings derived from participants' comments, we discuss some insights for the developing better MER systems in the following.

From the **instrument** theme, as participants distinguished between instrumentation and were impacted in an emotional sense by instrumental roles and interactions within the performance, it indicates that using separate instrumental tracks or combinations of them for building music emotion recognition models

might help to improve the prediction accuracy, comparing to modelling emotion directly from the mixed/mastered audio. Previous work by [28] has achieved a better emotion recognition results using multi-track audio of a small group of rock music. With more multi-track datasets [6, 19] open to public nowadays, this is an interesting avenue to explore further. Also, as people associated their emotion judgements to specific patterns of *instrument interaction*, a better detection of numbers of instruments playing at a given time, a better understanding of long-term interaction between instruments as well as the role of each instrument through the audio analysis may benefit emotion prediction. From the **structure** theme, as "repetition in music" has been reported to influence participants' emotion judgement such as building up emotion, being able to detect repetitions from music may also benefit MER. From the **expression** theme, as people have associated emotion judgements with specific music embellishments, it would help to incorporate the automatic detection of vibrato or other music ornaments into building MER systems especially for time-varying music emotion recognition. Recent advances in the MIR field on playing technique detection may provide such opportunities, such as works of detection of vibrato in violin and erhu [20, 34], arpeggios in multiple instruments [3], pedalling in piano [21] and representative playing techniques in guitar [32, 9] and bamboo flute [33]. Moreover, finally the **visual cues** theme indicated that dynamics of emotional perception in live performance could be a multimodal phenomenon, and multimodal emotion sensing using computer vision [23] and audio can also be promising in the future design of music emotion studies.

4 Conclusion

Understanding how music affects listeners perception of emotion facilitates creating fair and unbiased music information retrieval systems. In this paper, we examined the time-varying music emotion perception from the participants in a complementary way: The collection of time-varying emotion ratings enabled a quantitative measure of emotion responses and retroactive rating reflection; while explanations from participants helped to highlight the reasons behind such emotion judgements. However, we did not give an exhaustive answer regarding listener-informed features but present the current state and experimental data that have been collected so far within this ongoing project. In the future work, we plan to re-conduct the thematic analysis with more coders to increase the validity and reliability of the results. We also plan to investigate the individual differences on time-varying music emotion perception involving music expertise and demographic information, as well as to investigate the reasons behind the varied levels of agreement in perceived emotion agreement over the performance. As one of the most important issue in MIR tasks is the identification of appropriate and well-functioning features, our current findings of listener-informed music features underpin the previous emotion studies, in addition, the identification of less employed music features such as instrumentation and ornaments also generate some insight for the improvement of MER systems.

References

1. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PloS one **12**(3), e0173392 (2017)
2. Aucouturier, J.J., Bigand, E.: Mel cepstrum & ann ova: The difficult dialog between mir and music cognition. In: ISMIR. pp. 397–402 (2012)
3. Barbancho, I., Tzanetakis, G., Barbancho, A.M., Tardón, L.J.: Discrimination between ascending/descending pitch arpeggios. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(11), 2194–2203 (2018)
4. Barthet, M., Depalle, P., Kronland-Martinet, R., Ystad, S.: Acoustical correlates of timbre and expressiveness in clarinet performance. Music perception: An interdisciplinary journal **28**(2), 135–154 (2010)
5. Barthet, M., Fazekas, G., Sandler, M.: Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. Proc. CMMR pp. 492–507 (2012)
6. Bittner, R.M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., Bello, J.P.: Medleydb: A multitrack dataset for annotation-intensive mir research. In: ISMIR. pp. 155–160 (2014)
7. Bradley, M.M., Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings. Tech. rep., Citeseer (1999)
8. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative research in psychology **3**(2), 77–101 (2006)
9. Chen, Y.P., Su, L., Yang, Y.H., et al.: Electric guitar playing technique detection in real-world recording based on f0 sequence pattern recognition. In: ISMIR. pp. 708–714 (2015)
10. Cowie, R., McKeown, G., Douglas-Cowie, E.: Tracing emotion: an overview. International Journal of Synthetic Emotions (IJSE) **3**(1), 1–17 (2012)
11. Eerola, T., Vuoskoski, J.K.: A review of music and emotion studies: approaches, emotion models, and stimuli. Music Perception: An Interdisciplinary Journal **30**(3), 307–340 (2013)
12. Fazekas, G., Barthet, M., Sandler, M.B.: The mood conductor system: Audience and performer interaction using mobile technology and emotion cues. In: 10th International Symposium on Computer Music Multidisciplinary Research (CMMR'13). pp. 15–18 (2013)
13. Flexer, A., Grill, T.: The problem of limited inter-rater agreement in modelling music similarity. Journal of new music research **45**(3), 239–251 (2016)
14. Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., Elowsson, A.: Using listener-based perceptual features as intermediate representations in music information retrieval. The Journal of the Acoustical Society of America **136**(4), 1951–1963 (2014)
15. Gabrielsson, A., Lindström, E.: The role of structure in the musical expression of emotions. Handbook of music and emotion: Theory, research, applications **367400** (2010)
16. Hiraga, R., Matsuda, N.: Graphical expression of the mood of music. In: Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on. vol. 3, pp. 2035–2038. IEEE (2004)
17. Imbrasaitė, V., Baltrušaitis, T., Robinson, P.: Emotion tracking in music using continuous conditional random fields and relative feature representation. In: Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on. pp. 1–6. IEEE (2013)

18. Juslin, P.N., Lindström, E.: Musical expression of emotions: Modelling listeners' judgements of composed and performed features. Music Analysis **29**(1-3), 334–364 (2010)
19. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Transactions on Multimedia **21**(2), 522–535 (2019)
20. Li, P.C., Su, L., Yang, Y.H., Su, A.W., et al.: Analysis of expressive musical terms in violin using score-informed and expression-based audio features. In: ISMIR. pp. 809–815 (2015)
21. Liang, B., Fazekas, G., Sandler, M.: Piano sustain-pedal detection using convolutional neural networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 241–245. IEEE (2019)
22. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on audio, speech, and language processing **14**(1), 5–18 (2006)
23. Mou, W., Gunes, H., Patras, I.: Alone versus in-a-group: A multi-modal framework for automatic affect recognition. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **15**(2), 47 (2019)
24. Müllensiefen, D., Gingras, B., Stewart, L., Musil, J.J.: Goldsmiths musical sophistication index (gold-msi) v1. 0: Technical report and documentation revision 0.3. London: Goldsmiths, University of London. (2013)
25. Rusell, J.: A circumplex model of affect. Personality and Social Psychology pp. 1161–1178 (1980)
26. Schmidt, E.M., Kim, Y.E.: Modeling musical emotion dynamics with conditional random fields. In: ISMIR. pp. 777–782 (2011)
27. Schubert, E.: Modeling perceived emotion with continuous musical features. Music Perception: An Interdisciplinary Journal **21**(4), 561–585 (2004)
28. Scott, J., Schmidt, E.M., Prockup, M., Morton, B., Kim, Y.E.: Predicting time-varying musical emotion distributions from multi-track audio. CMMR **6**, 8 (2012)
29. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979)
30. Soleymani, M., Aljanaki, A., Yang, Y.H., Caro, M.N., Eyben, F., Markov, K., Schuller, B.W., Veltkamp, R., Weninger, F., Wiering, F.: Emotional analysis of music: A comparison of methods. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1161–1164. ACM (2014)
31. Sturm, B.L.: Evaluating music emotion recognition: Lessons from music genre recognition? In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6. IEEE (2013)
32. Su, L., Yu, L.F., Yang, Y.H.: Sparse cepstral, phase codes for guitar playing technique classification. In: ISMIR. pp. 9–14 (2014)
33. Wang, C., Benetos, E., Lostanlen, X., Chew, E.: Adaptive time–frequency scattering for periodic modulation recognition in music signals. In: ISMIR (2019)
34. Yang, L., Rajab, K.Z., Chew, E.: The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation. Journal of Mathematics and Music **11**(1), 42–60 (2017)
35. Yang, S., Barthet, M., Chew, E.: Multi-scale analysis of agreement levels in perceived emotion ratings during live performance. In: Extended abstracts for the Late-Breaking Demo Session of ISMIR (2017)
36. Yang, Y.H., Liu, J.Y.: Quantitative study of music listening behavior in a social and affective context. IEEE Transactions on Multimedia **15**(6), 1304–1315 (2013)

Towards Deep Learning Strategies for Transcribing Electroacoustic Music

Matthias Nowakowski[1], Christof Weiß[2], and Jakob Abeßer[3]

[1] Media Informatics, University of Applied Sciences, Düsseldorf
[2] International Audio Laboratories Erlangen
[3] Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau
matthias.nowakowski@gmail.com

Abstract. Electroacoustic music is experienced primarily through hearing, as it is not usually based on a prescriptive score. For the analysis of such pieces, transcriptions are sometimes created to illustrate events and processes graphically in a readily comprehensible way. These are usually based on the spectrogram of the recording. Although transcriptions are often time-consuming, they provide a useful starting point for any person who has interest in a work. Deep learning algorithms, which learn to recognize characteristic spectral patterns using supervised learning, represent a promising direction of research to automatize this task.
This paper investigates and explores the labeling of sound objects in electroacoustic music recordings. We test several neural network architectures that enable classification using musicological and signal processing methods. We also show future perspectives how our results can be improved and how they can be applied to a new gradient-based visualization approach.

Keywords: electroacoustic music, acousmatic music, transcription, deep learning

1 Introduction

Scientific discourse is based on an intersubjectively accessible object and its representations. Musicology usually treats music as sound or score. Especially studying electroacoustic music, the approach must always be hearing, since the peculiarity of this type of music is that the sound is fixed and not its prescriptions. Sound material is either synthetically produced or consists of electronically processed recordings [1,2]. Thereby, timbre and its temporal progression became important structural elements, in contrast to harmony or metre [3]. In order to make this music comparable, transcriptions are created in practice, which are mostly colorations and linguistic annotations of spectrograms [4].

There are not many attempts to automate this process for electroacoustic music. Especially the lack of a uniform nomenclature to describe sound objects is an issue to be discussed. Analyzing such properties with signal processing has been addressed sparsely [5].

In recent years neural networks have shown the best results for tasks such as genre detection, chord recognition, speech recognition, etc., in the analysis of tonal music [6]. In particular end-to-end visualization techniques, as well as better scalability on dataset size make this state-of-the-art technology interesting for exploring this task. The few papers which dealt with machine learning and electroacoustic music, either treat this subject superficially, or some advantages where not yet widely accessible at the time of publication [7].

Such an endeavor can not only be fruitful for speeding up musicological discourse processes. Transcriptions are also an communicative device which can be used exploratively (find morphologies which were not heard before, reveal macroforms) or explanatively (backing up individual transcriptions by technical means) and can so enhance accessibility.

2 Previous Work

In the past, common Music Information Retrieval (MIR) techniques where used to analyze and visualize certain features within elecotracoustic pieces and are also implemented as software [8,9]. Also further features which capture style-relevant properties could be used for electroacoustic music as well [3, 5]. But interest in using expensive machine learning algorithms is rather low, although deep learning approaches provide state-of-the-art results in different music related tasks [6].

In a recent study, Collins made thorough analyses of electroacoustic pieces after the release of a large, online available corpus.[4] He used fully connected networks on previously extracted features to estimate publication years [10], but was not pursuing this approach in subsequent publications [11], since k-Nearest-Neighbor algorithm outperformed neural networks in accuracy.

Klien et al. discuss the use of machine learning more critically from an aethetic standpoint [12]. In their view, fully automated annotations are not able to overcome the semantic gap between the signal and the meaning, since electroacoustic (or acousmatic) music tries do defy the definition of music itself. Any algorithm used for analysis therefore should not attempt to "understand" music. Despite their position, we agree that a human annotator is needed to make reasonable assertions about musical structure. In contrast, one could consider the complexity of the task in particular to be suitable for a deep learning approach.

Using interactive hierarchical clustering, Guluni et al. [13,14] let experts categorize sound events of different timbres of a synthetic data set giving this results into a feedback loop. The authors use a Support Vector Machine (SVM) classifier, which is fed with the feature coefficients. Both monophonic and polyphonic compositions had results with F-measures ≥ 0.85.

Given a sound scene (being musical or not) in which sounds change characteristics according to auditory distance or distortion, form coherent textures with other sounds, or split from them gradually, it might be helpful to view

[4] `http://www.ubu.com/sound/electronic.html`

this task as one lying close sound event detection and sound scene analysis. One of the main challenges is high intraclass variance due to the quality of of sounds which may be not clearly separated from one another and appear in multisource environments [15]. As shown in the regularly held "Detection and Classification of Acoustic Scenes and Events" (DCASE) challenges best results are produced by neural networks which are trained on some type of a spectrogram representation [16, 17]. From this point of view, deep learning seems like a viable approach for elecotracoustic music. The main problem is to develop a suitable set of labels, which show satisfactory results in classification tasks before temporal sound event detection can even take place. Using deep learning also gives the possibility to employ methods for interpreting deep neural networks to generate visualizations [18]. This might show what portions of a spectrogram were actually important while training on sound event classification and so gives visual cues to examine the dataset.

3 Data & Annotation

In this section, we describe the creation and annotation of a dataset of electroacoustic music excerpts. Since there is no consistent or commonly used nomenclature for categorizing sound objects, analysis frameworks can help to develop labels which could be used and understood by musicological experts. For better comparison with previous approaches, we use label names in accordance with commonly used features which are used in other classification algorithms. Although there are some historically relevant frameworks like Pierre Schaeffer's *Typomorphology* [19], for the labels used here we draw our inspiration from Denis Smalley's *Spectromorphology* [20]. He developed this framework as a tool for describing sound-shapes, i. e. the unit of spectral content and their development in time, based on aural perception. Adopting this viewpoint can be helpful to identify such sound-shapes in a spectrogram which is our base baseline feature.

We chose the five descriptors *flux*, *spread*, *noise*, *density*, and *realness* as attributes to describe both the static and dynamic aspects of spectromorphological expectation. For each attribute, the extreme values (0/1) represent poles within a description space.

Flux 0: Stationary; 1: Fluctuating
Spread 0: Narrow spectral range; 1: Wide spectral range
Noise 0: Harmonic; 1: White Noise
Density 0: Single event; 1: Multiple events (uncountable)
Realness 0: Synthetic; 1: Real world sound source

Flux and *density* were selected to reflect the development of a sound event over time. In contrast, *spread* and *noise* describe static sound characteristics. All attributes can be combined to form a label set to provide more complex descriptions of sound events. Each attribute in a label is represented by its initial letter. We obtain 32 possible classes from all combinations of the five attributes. For instance, *f0s0n1d1r0* represents a stationary and narrow sound,

which is very noisy, has a high density and a synthetic sound characteristic. As an analogy, one could think of a pass filtered noise band. Similarly, we can define four separate classes by combining only two attributes. An example of such an annotation w.r.t. *spread* and *noise* could be *s0n1*, which defines a filtered noise-like sound without specifying its temporal characteristics. On the one hand, this way of choosing attributes to form new label sets allows to refine classes during the annotation process. On the other hand, a binary attribution can lead to fuzzy class boundaries, so that event labeling becomes imprecise. For instance labeling *density* of a drum roll may diverge into labeling each event itself or the whole texture depending on the event frequency. While each event is probably static, the texture could gain fluctuating characteristics due to playing style or instrument manipulation, like a glissando on a timpani. Therefore, during the annotation process, we focused on sound objects, which can be clearly defined by the selected attributes. Silence can not be reflected by these attributes.

The compiled dataset consist of excerpts of 29 electroacoustic pieces from the 1960s to the early 2000s. Longer pieces were cut off at the 5 minutes mark, while whole pieces were taken if they were shorter or slightly longer than 5 Minutes. This adds up to a total duration of 2.5 hours.

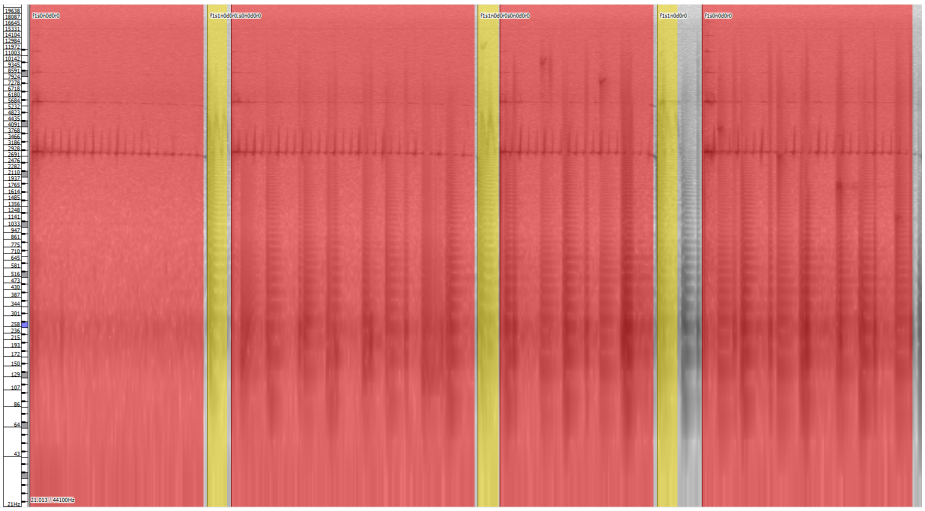


Fig. 1: Example annotation in the Log-Spectrogram of the first 12 seconds of the Movement *Polyrythmie* in *Vibrations Composés* by François Bayle. Red boxes represent recurring/ similar events annotated as *f1s0n0d0r0*, while yellow boxes represent recurring/ similar events annotated as *f1s1n0d0r0*. Although more events can be seen (e.g. the dark colored band in the lower half of the spectrogram), all frequencies are used as an input feature for each annotated event as indicated by the boxes.

Each relevant sound event was then annotated manually including the attack and end of release. Since almost all recordings are polyphonic, some sound

events may appear in multiple longer events. This leads to a total of 3.7 hours of extracted material. 3016 separate events were annotated ranging from 0.05 seconds to 4.5 minutes. We enlarged the dataset using data augmentation techniques. To this end we applied mild pitch shifting using step sizes between minus and plus 2 semitones in order to not distort the spectral characteristics. In total, the dataset contains 18.5 hours of segmented and augmented material.

Since some classes are stronger represented than others, all the extracted events were resampled to the mean duration of all 32 classes. Resampling on these classes also scales to all other possible label sets. Longer classes were shortened by randomly deleting events. In shorter classes events were duplicated in turn with slight random differences in the signal.

In our experiment, we repeat a random dataset split into training, validation, and test set using a split ratio of 60%, 20 %, 20 % three times by number of events. We ensure that the events in the three evaluation sets come from unique recordings.

4 Experiments

In this paper we focus on reporting results from configurations made with a 4-class label set consisting of the attributes *spread* and *noise* to investigate the impact of a combination of static spectral attributes at first. By using a label set consisting of two attributes we reduce chances of wrong labeling and have a more manageable number of attributes to compare. For all experiments the following labels were used: *s0n0*, *s0n1*, *s1n0*, *s1n1*. The deep learning architectures were implemented using the Keras framework[5], wheras feature extraction algorithms were implemented after [21] or used directly through the librosa library[6]. For our experiments we have focused primarily on the performance of the classification.

4.1 Metrics

Test F1-score This measure is used to score the overall classification of a sound event. The F1-score equals the harmonic mean between precision and recall. If classification is done on multiple patches of an event, the mean is computed.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1}$$

Training validation accuracy difference (ΔAcc) The accuracy for both training and validation set is computed during the training itself. The goal is to have training and validation accuracy as close as possible. A lower value for ΔAcc means less overfitting.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\Delta Acc = |Acc_{train} - Acc_{val}| \tag{3}$$

[5] `https://keras.io/`

[6] `https://librosa.github.io/librosa/`

All metrics are averaged over the number of folds in the respective configuration. Early stopping on validation accuracy was employed for all the deep learning experiments after 20 epochs due to usually very short training (one digit numbers of epochs) when using early stopping on the validation loss.

4.2 Baseline

Because we use a self-developed annotation scheme and data set for this paper, the definition of a baseline performance is challenging. For four classes, we expect the random baseline to be $P(4) = 0.25$. To compare our results to classical approaches, we used a Random Forest classifier with 100 estimators. For this we used "spectral variation" (also called "spectral flux") for *flux* and "spectral spread" for *spread* as described in [21]. *Noise* is described by the absence of the estimated harmonicity according to a detected fundamental frequency. As a feature for *density* we used the residual part of an HRPSS [22]. In lack of a descriptor for *realness* we just used the mean of the mel-spectrogram. All coefficients were computed on the complete event by averaging over time.

Using this classifier also gives the possibility to determine importances of all feature coefficients. Although using five features we expect higher importances for the ones corresponding to our attributes.

4.3 Convolutional Neural Network (CNN)

We now want discuss the CNN architectures used for our experiments. As input representation, we chose mel-spectrograms to keep the aspect ratio of sound-shapes independent of vertical position. Time-frequency transformations were made from wav-files with a sample rate of 22050 Hz using a window size of 1024 frames, a hop length of 512 frames and a vertical resolution of 96 bins. Each spectrogram was then cut into patches of 100 frames (around 2 seconds) with 50% overlap of each subsequent patch. Shorter events were padded by adding low positive noise (up to 10^{-4}) on the spectrogram to reach the desired minimum patch length. Values where scaled by appling zero-mean unit variance normalization.

For the CNN configurations we chose to use a shallow version of the VGG-Network with 7 Layers [23]. Here, we compare architectures using 2D and 1D convolutional layers. Each architecture consists of two convolutional layers with 32 kernels, followed by two layers with 64 kernels. After each convolution, we applied batch normalization and max pooling. Classification was done with following fully connected layers using one dense layer with 512 nodes as well as a final dense layer with four output nodes and softmax activation functions. We used dropout (0.25), L2 kernel regularization (0.01) after the first dense layer, as well as adding Gaussian noise (0.1) on the input in order to regularize the model and reduce overfitting. For the 2D convolutional architecture, we use 3x3 kernels and apply global average pooling before the FCN. Accordingly, we use convolution kernels of size 3 for the 1D architectures before the CNN output is flattened and forwarded to the fully connected layers. Both architectures were

chosen due to their different approach on computing features of the given input. While the 2D convolution is able to detect position invariant patterns, 1D convolution focuses on local patterns in a sequence.

4.4 Convolutional Recurrent Neural Network (CRNN)

For the CRNN, a bi-directional Gated Recurrent Unit (GRU) layer with 32 units was added after the CNN processing for temporal modeling. We chose GRU over Long Short-Term Memory (LSTM) units, because of faster model convergence while showing similar performance to LSTM [24]. The first advantage of using CRNN over CNN alone is that this architecture can better model long-term dependencies. Secondly, such a network architecture can be trained with variable-length input data, while CNNs require fixed-size input. However, we observed a strong model overfitting when training CRNNs on full spectrograms of a piece. Therefore, we will focus on reporting results from the CNN model trained with a fixed input tensor size first. Then, we evaluate, whether the classification results improve if the CRNN is instead initialized with the trained parameters from a CNN model.

5 Results

	Architecture	*F1*	ΔAcc
Random Baseline	-	0.25	-
Shallow Classifier	Random Forest, 100 Estimators	0.27	-
CNN 2D	2 x Conv 2D 32, 2 x Conv 2D 64, 512 FCN	0.335	0.152
CNN 1D	2 x Conv 1D 32, 2 x Conv 1D 64, 512 FCN	0.315	0.207
CRNN 2D	2 x Conv 2D 32, 2 x Conv 2D 64, Bidirectional GRU 32, 512 FCN	0.362	0.112
CRNN 1D	2 x Conv 1D 32, 2 x Conv 1D 64, Bidirectional GRU 32, 512 FCN	0.385	0.022

Table 1: Results of the 4-class classification experiments

Feature importances for the baseline experiment show slight tendency towards the mel spectrogram with 0.22, while the noise feature had the least impact on the classification with 0.16. Importances of the remaining features are balanced.

Comparing our approach with the baseline performance, deep learning improves classification results to some extent. Only using CNN layers for computation, 2D convolution gave best results with a F1-value of 0.335 over 1D

flux	*spread*	*noise*	*density*	*mel*
0.21	0.21	0.16	0.21	0.22

Table 2: Feature importances for the baseline experiment (Random Forest classifier)

convolution with 0.315. Accuracy differences are still relatively high so that we can observe some amount of overfitting. But taking each fold into account, standard deviation over all accuracy differences in CNN 2D with 0.14 is relatively high as compared to CNN 1D with 0.089. For each fold, a higher ΔAcc usually correlates with higher numbers of training epochs, being a maximum of 98 epochs for CNN 2D and 38 for CNN 1D (Table 3).

	CNN 2D		*CNN 1D*	
	Epochs	ΔAcc	*Epochs*	ΔAcc
Fold 1	3	0.123	16	0.265
Fold 2	98	0.361	38	0.275
Fold 3	1	0.028	3	0.081
Standard Dev.	-	0.14	-	0.089

Table 3: ΔAcc for each fold in CNN 2D and CNN 1D

Adding the GRUs to the architectures and initializing weights with the aforementioned models results in improved classification results. Especially, CRNN 1D outperforms all experiments with a F1-value of 0.385, increasing by 0.07, whereas the F1-value for the CRNN 2D increases just by 0.027 to 0.362. ΔAcc decreases for both experiments. For CRNN 1D by 0.185 and 0.04 for CRNN 2D. Also the maximum training time decreased for for both experiments being it 12 epochs for CRNN 2D and 14 epochs for CRNN 1D. Overall, a high F1-value correlates with a lower ΔAcc pointing to less overfitting.

6 Discussion

In this paper we presented a musicologically informed approach for sound object classification of electroacoustic music. Using deep learning, we could show that some improvement could be achieved by architectures more sensitive to sequential data, which can facilitate classifying data as described by our morphological labels. Despite reducing the label space, feature importances for the selected attributes *spread* and *noise* do not have any significant impact on the classification in the baseline experiment. *Noise* even had the lowest importance. This shows that the semantic implications of the chosen attributes are not transferable to common descriptors that easily, so that they require more complex feature sets.

Although we can see that CNN 2D had better classification results, using CNN 1D resulted in constant generalization throughout the folds (Table 3). This indicates that features depend less on position invariant sound-shapes, but on the vertical integrity of the spectrum, or rather more spectral context than previously expected. One approach to validate this in future experiments is to apply horizontally-shaped CNN filters instead of symmetrical ones to incorporate a larger spectro-temporal context. The importance of temporal succession over the isolated position in the spectrum is then pronounced by the improved scores using CRNN.

Since there are still many questions and many configurations to be tested, this paper is merely a suggestion and baseline on further investigation into this field and even this approach can be still evaluated on more or different attributes, features and architectures. In general, results lagged far behind our initial hopes, which can be attributed to our more explorative approach of this problem. At this point our results are not satisfactory and the performance is not far above the random baseline so we decided not to present a musicological analysis. With regard to the transcription of sound objects, the outcome of such experiments can be used for visualization, using e.g. gradient-based network analysis algorithms. These show portions of the spectrogram, which were relevant for the classification. We suspect these means to be helpful detecting and displaying sound objects in the spectrogram. For our purposes we tried layer-wise relevance propagation (LRP) [18] which resulted in relevancy maps which are merely hints to what the network actually learns. But the classification scores are quite low so that mappings at this point are mostly inconclusive and still have to be evaluated.

While developing and evaluating the experiments we noticed some issues which we want to address and propose some future solutions. This could help develop next steps more carefully.

Labeling approach During the annotation procedure, only one person familiar with electroacoustic music worked on the labels. Thus, no validation could be made. To reduce the bias of the annotator, a group of people could cross-check their annotations. In different classification tasks, as e.g. discriminating cats and dogs, we can expect human experts to classify all samples right. Such bias values for a complex task like the one presented here do not exist so that cross-checking labels could also be a basis for more research. In addition, one could think about continuous values for annotating attributes. This could lead to embeddings of such sound objects which might help constructing new labels or label families.

Dataset size The dataset used in this paper is relatively small. Therefore, more elaborate transfer learning techniques [25, 26] could be employed following the assumption that suitable low level features can be extracted from different music related tasks [27] or even different domains such as images [28]. Beside feature transfer, one could also apply multitask learning if labels for both source and target task are available [25]. The main idea is to train source and target task simultaneously, provided both tasks are similar, to

extract shared features. Even in the case of negative transfer, analyzing predicted targets could help in investigating the most helpful features. For a source task e.g. electroacoustic sound sources can be used such as *electronic*, *environment*, *instrument*, or *singing* while the target task labels remain morphological attributes.

Studio effect To validate the chosen labels we wanted to see if unsupervised machine learning techniques could help to figure out if some consistency in the data points can be found beyond our chosen descriptors. To this end, over 160 features were extracted for all segments according to [29] which where used in electroacoustic music related tasks in [13, 14]. Using t-SNE [30] for dimensionality reduction and DBSCAN for clustering we found that the studio effect, the impact of the production conditions on the feature extraction, had large impact on the clustering. In Fig. 2, each colored cluster consists of most segments of one piece used in the dataset, except blue which consists of segments of all other pieces. Grey transparent points are considered to be noisy data points by the clustering algorithm. To avoid this effect, we would need a more uniform dataset composed by some experts especially for that task, such as in [13,14][7], to avoid this effect. This is not a statement about, if our descriptors do work, or not. It rather is an example of a problem we came across and which we have to face when designing a dataset using samples from a field of music which is highly dependent on its medium.

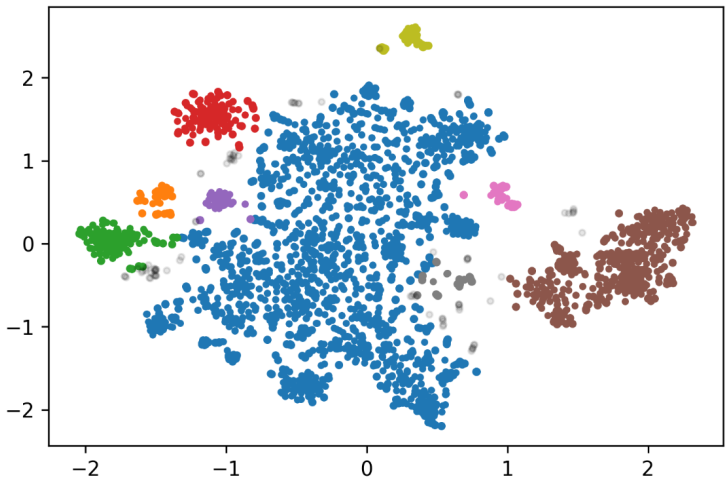


Fig. 2: Visualization using t-SNE for dimensionality reduction and DBSCAN for clustering. Each colored cluster consists of most segments of one piece used in the dataset. The blue cluster consists of segments of all other pieces.

[7] We have requested this dataset, but unfortunately it was no longer provided by the creators.

Acknowledgements. This work has been supported by the German Research Foundation (AB 675/2-1, MU 2686/11-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

References

1. Stroh, W.M.: Elektronische Musik. Handbuch der musikalischen Terminologie 2, Steiner-Verlag, Stuttgart (1972)
2. Beiche, M.: Musique concrète. Handbuch der musikalischen Terminologie 4, Steiner-Verlag Stuttgart, (1994)
3. Weiß, C., Müller M.: Quantifying and Visualizing Tonal Complexity. In: Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM), pp. 184–187, Berlin (2014)
4. Erbe M.: Klänge schreiben: die Transkriptionsproblematik elektroakustischer Musik. Apfel, Vienna (2009)
5. López-Serrano, P., Dittmar, C., Müller M.: Mid-Level Audio Features Based on Cascaded Harmonic-Residual-Percussive Separation. In: Proceedings of the Audio Engineering Society AES Conference on Semantic Audio, Erlangen (2017)
6. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang S.-Y., Sainath, T.: Deep Learning for Audio Signal Processing. IEEE Journal of Selected Topics in Signal Processing 14/8, 1–14 (2019)
7. Pons, J.: Neural Networks for Music: A Journey Through Its History, `https://towardsdatascience.com/neural-networks-for-music-a-journey-through-its-history-91f93c3459fb` (2018)
8. Couprie, P.: Methods and Tools for Transcribing Electroacoustic Music. In: International Conference on Technologies for Music Notation and Representation – TENOR'18, pp. 7-16, Montral (2018).
9. Park, T.H., Li, Z., Wu, W.: Easy Does It: The Electro-Acoustic Music Analysis Toolbox. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 693–698, Kobe (2009)
10. Collins, N.: The UbuWeb Electronic Music Corpus: An MIR investigation of a historical database. Organised Sound 20/1, pp. 122–134 (2015)
11. Collins, N., Manning, P., Tarsitani, S.: A New Curated Corpus of Historical Electronic Music: Collation, Data and Research Findings. Transactions of the International Society for Music Information Retrieval 1/1, pp. 34–55 (2018)
12. Klien, V., Grill, T., Flexer, A.: On Automated Annotation of Acousmatic Music. Journal of New Music Research 41/2, 153–173 (2012).
13. Gulluni, S., Essid, S., Buisson, O., Richard, G.: An Interactive System for Electro-Acoustic Music Analysis. In: 12th International Society for Music Information Retrieval Conference (ISMIR 2011), pp. 145–150, Miami (2011)
14. Gulluni, S., Essid, S., Buisson, O., Richard, G.: Interactive Classification of Sound Objects for Polyphonic Electro-Acoustic Music Annotation. AES 42nd International Conference, Ilmenau (2011)
15. Virtanen, T., Plumbley, M.D., Ellis, D.P.W.: Computational analysis of sound scenes and events. Springer Verlag, Cham (2018)

16. Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., Virtanen, T.: DCASE 2017 Challenge setup: Tasks, datasets and baseline system. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events (2017)
17. Adavanne, S., Virtanen, T.: A Report on Sound Event Detection with Different Binaural Features. DCASE2017 Challenge (2017)
18. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek W., Müller, K.-R., Dähne, S., Kindermans, P.-J.: iNNvestigate neural networks! CoRR (2018)
19. Thoresen, L., Hedman, A.: Spectromorphological Analysis of Sound Objects: An Adaptation of Pierre Schaeffer's Typomorphology. Organised Sound 12/2, pp. 129–141, Cambridge (2007)
20. Smalley, D.: Spectromorphology: Explaining Sound-shapes. Organised Sound 2/2, pp. 107–126, Cambridge (1997)
21. Peeters, G.: A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project, `http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf` (2004)
22. Drieger, J., Müller M., Disch S.: Extending Harmonic-Percussive Separation of Audio Signals. In: Retrieval Conference (ISMIR 2014), pp. 611–616, Taipei (2014)
23. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. ILCR (2015)
24. Chung, J., Gülçehre, C., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. NIPS 2014 Deep Learning and Representation Learning Workshop (2014)
25. Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22/10, pp. 1345–1359 (2010)
26. Torrey, L., Shavlik, J.: Transfer Learning. In: Handbook of Research on Machine Learning ,Algorithms, Methods, and Techniques, pp. 242–264, IGI-Global (2009)
27. Choi, K., Fazekas, G., Sandler, M.B., Cho, K.: Transfer Learning for Music Classification and Regression Tasks. In: Proceedings of the 18th ISMIR Conference, pp. 141–149, Suzhou (2017)
28. Grzywczak, D., Gwardys, G.: Deep Image Features in Music Information Retrieval. Intl. Journal of Electronics and Telecommunications 60/4, 321–326 (2014)
29. Essid, S., Richard, G., David, B.: Musical Instrument Recognition by Pairwise Classification Strategies. IEEE Transactions on Audio, Speech, and Language Processing 14/4, 1401–1412 (2006)
30. van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)

Improvisation and Environment

Christophe Charles [1]

[1] Musashino Art University, Tokyo 187-8505, Japan
charles@musabi.ac.jp

Abstract. To improvise is to "create and perform spontaneously or without preparation" and without being able to predict the outcome. One often distinguishes composed music and improvised music, but as it is impossible to write everything, and all music necessarily has a part of improvisation, we can consider that all music, beyond a certain degree of (non) writing, is improvised. John Cage has long rejected improvisation which would often favor subjective choice and intention, but he eventually found a way to accept the idea of "structural improvisation", through a re-definition of space and time in music. Other composers, such as Christian Wolff, John Russell, Pauline Oliveros or Kosugi Takehisa, have also developed different ways to improvise by questioning the self in relation to its environment. With information technologies, the composer is now able to be a performer and a listener at the same time. Composing-performing-listening can be considered as the activity of exploring what is happening in real-time, not only between the subject and its environment, but also between different levels of consciousness of the subject or between subjects. This "in-between" is called "aida" by Kimura Bin. The above forms of improvisation might be useful to explore the "aida".

Keywords: Improvisation, Intentionality, Environment.

1 Improvised Music and Written Music

To improvise is to "create and perform spontaneously or without preparation" [1]. It is borrowed from the Latin improvises: "which arrives unexpectedly". In Japanese and Chinese, "improvisation" translates as sokkyô [即興]: "act immediately" (to compose for example a poem or music). I therefore propose to approach the idea of improvised music as a practice of simultaneous and immediate composition and execution, without being able to predict the outcome.

One often distinguishes composed music and improvised music, by posing that the composed music is written, and the improvised music is not written. Now it is impossible to write everything, and all music necessarily has a part of improvisation, which means that beyond a certain degree of (non) writing, all music is improvised. Rules and conventions, written or oral, which order and control improvisation, are defined to guarantee the authenticity of musical composition and performance. These rules allow more or less freedom to the performer's spontaneity, in temporal and spatial choices, choice of volume, timbres and pitches, techniques, ideas which relate to aesthetics, ethics, philosophy, sociology, politics, marketing, etc. "Written" music having been fixed after many reflections and corrections, is often considered more serious than improvised music, which is unpredictable and therefore difficult to judge. I will first examine the ideas of some composers - performers who have thought about the difference or rather the balance between composition and improvisation, and will try from there to summarize my own concerns and activities.

2 Intentionality and Non-intentionality

In 1949, Cage defined the four basic elements of a musical composition: "Structure in music is its divisibility into successive parts from phrases to long sections. Form is content, the continuity. Method is the means of controlling the continuity from note to note. The material of music is sound and silence. Integrating these is composing." [2] Cage then uses the term "experimental" for "an action the outcome of which is not foreseen," [3] and is particularly interested in the idea of indeterminacy in music and the other arts, thinking about how we can or cannot control the degree of improvisation and freedom when performing a composition, to produce a result that is really unexpected. Paradoxically, a composition can be highly indeterminate if fully controlled: the performer of 4'33" has only two things to do, look at his watch and keep quiet.

The quest for the unexpected implies the recognition of the "reality" of sounds: "sounds are to come into their own, rather than being exploited to express sentiments or ideas of order" [4]. It was also Stravinsky's opinion: "Music, by its essence, is powerless to express anything: a feeling, an attitude, a psychological state, a phenomenon of nature, and so on. The expression has never been the immanent property of music. [5] Clément Rosset explains in another way the nature of music and its relation to reality: "Music is a creation of reality in its raw state, without comment or reply; and the only case where the real is presented as such. This is for a very simple reason: music does not imitate, it exhausts its reality in its only production, such as the *ens realissimum* - the supreme reality - by which metaphysicians characterize the essence, it is model to anything but is itself modeled on nothing [...] Music is an advanced point of reality, something like the "eminence" of reality, offered to perception as a sort of preview of reality." [6] Cage wants thus "both in sound, sight, graphic arts and music and so forth, to experience each sound and each thing we see as itself, rather than as representative of something else." [7] The interpreter should not to try to give a meaning that exists only in his imagination.

It was after a long process of reflection, especially from the time of his meeting with Suzuki Daisetsu in the late 1940s, that he was able to define a discipline to overcome his tastes and memories: "The ego has the capacity to cut itself off from the rest of mind, or to flow with it. It does that by developing likes and dislikes, taste and memory, and if you do what Zen wants you to do, that is, get free of your taste and memory and likes and dislikes, then you have to discipline yourself. My discipline is that of the I-Ching, shifting my responsibility from making choices to asking questions, and getting the answers, by means of the ancient coin tossing method of the I-Ching." [8] Cage explains to Roger Reynolds: "What is concerned in each case is the ego that is being controlled in such a way that it will not act as a barrier to its experience, but that it itself will change. In other words. discipline is implicit in the idea of a question as opposed to a choice." [9] The refusal of choice removes the expression of one's tastes, prejudices and memory, and opens the possibilities. Cage has thus developed the idea of non-intentionality, that is, of composition without intention. However, just as it is necessary to make the choice of not making a choice, one must intend not to have any intention. In other words, we cannot avoid either choice or intention, but can nevertheless decide to what extent we make a choice and allow ourselves to have intentions. From the 1970s, Cage restores the possibility to use intention in the compositional process: "chance procedures are only one tool among many that Cage has used to pursue quite consistently a single goal throughout his career: the disciplined acceptance, in musical contexts, of that which had been previously rejected out-of-hand." [10]

This applies to composition, but what about performance that necessarily requires improvisation? Carl Dalhaus points out the danger of using ready-made formulas: "To avoid embarrassment or silence, the improviser must be able to rely on a repertoire of pictures available at any moment, on a ground of prefabricated elements that s/he

modifies and agency variously, but that s/he does not invent on the field. The idea that s/he abandons himself unreservedly to the favor or disfavor of chance is a fiction." [11]. Cage wants above all to avoid clichés, and in particular self-expression: "What I would like to find is an improvisation that is not descriptive of the performer, but is descriptive of what happens, and which is characterized by the absence of intention. It is at the point of spontaneity that the performer is most apt to have recourse to his memory. He is not apt to make a discovery spontaneously." [12]

But how is it possible to perform a composition in an unintentional way, especially when it is complex to play, and requires a seemingly intentional control of play? We could paraphrase here the famous word by Zhuang-zi: the musician must be able to play "without effort", such as the swimmer who moves freely in the whirlpools of a waterfall: "grew up with my nature, and let things come to completion with fate" [13]. Zhuang-zi insists here on the intelligence of the body, and shows through the stories of the butcher and of the swimmer that there are two forms of action, one necessary and spontaneous (the Celestial order), and the other intentional and conscious (the Human order). Commenting Jean-François Billeter's interpretation, Jean Levi explains that "the spontaneous act is superior to the intentional act because, mobilizing all the capacities that are in us and naturally bending to the exigencies of the environment and circumstances, it escapes the errors of the intellect, which is dependent on random guessing and constrained by all sorts of prejudices." [14].

But it is not given to anyone to swim or play "without effort". The task of the artist will be to start the process of rehabilitation of the improvisation. Like the wandering visitor in a Japanese garden, who is invited, according to the word of Yamaguchi Katsuhiro [15], to walk circularly [回遊], the interpreter of Cage's "Number Pieces" will have to improvise the temporal structure of the composition, without having to worry about questions of taste: "Being anywhere in the time is part of my notion of improvisation. [...] I would call it 'structural improvisation.' By that, I mean an improvisation that exists in a given period of time, and that time is divided into 'rooms.' The instruments, whatever they happen to be, are played in such a way as to make the difference between those rooms clear. But where the sound is in the room is not important. Or you could say it could go "in any direction"… [...] the performer, the improvisor, and the listener too are discovering the nature of the structure, [...] the means that one can use to clarify the structure." [16]

William Brooks notes that recourse to chance operations is always possible, but no longer necessary, if one posits that "that which is arrived at by choice is in no sense preferable to that arrived at by chance." Cage thus rehabilitates intention and choice, but leaves aside the "values", which are not part of the music: "values contain neither sound nor silence. Values concern interactions - in this case, between people and sounds - and such interactions are essentially social in character." Cage thus redefines music by getting rid of these values, which "contain neither sound nor silence. Values concern interactions - in this case, between people and sounds - and such interactions are essentially social in character". [17]

3 Other Approaches of Improvisation

Let us see how other composers who, having understood Cage's approach, have defined their own approach. For Christian Wolff, the performers must be able to react with each other. The composition is a system that allows to "reveal to both performers and listeners energy resources in them themselves of which they hadn't been aware, and put those energies intelligently to work." [18] The performer(s) must use their technique but also their imagination to configure their participation. All sounds are considered environmental, and the performers have to find out how to coordinate "their" sounds with external sounds, those of the environment or those of other

performers. Christian Wolff defines no basic rhythm, and there is no conductor. The system of the composition does not allow the performers to play in a predictable way, because the sound interventions are relatively short, and they are constantly on the alert by reacting to what is happening around them: "You can't lay out a whole map and know exactly the path you're going to go. This means that you may be at a certain point in a piece where you suddenly don't know what 's next until someone else tells you. That's the one thing. The other thing is that the other person may not know that they're telling you something. The point of that has to do with eliminating as much as possible total control by any one person. It's almost impossible to conduct my music, for example. Everybody has to conduct, not all at once, but they take turns. Or they do it unintentionally." [19]

Guitarist John Russell wants to explore all the sounds he can produce with his instrument, because any sound can be used as a musical material. A musician has a set of "filters", and is "constantly analyzing, revaluing and reorganizing these, while at the same time trying to find new ones to generate further possibilities; but it is in the act of improvising that 'quantum leaps' can occur. Indeed, sometimes the whole architecture crumbles, leaving nothing at all as a reference." The instrument itself is also considered as an environment which has its own voice. It must be honored, that is, not (only) mastered but listened to. The relationship between the musician and his instrument is constantly changing, as the instrument too reacts to changes in the environment. For Russell, "any attempt at universality always fails", and it is precisely the errors and mistakes that make it possible to find openings: "the attack of a note carry a wealth of information and the 'suffocation', 'scratched' or 'aphasic' can be the results of exploring this. It can also come from a desire for imprecision or ambiguity which is a very useful tool in improvisation and indeed in certain types of composed music." [20]

Pauline Oliveros speaks of improvisation as a "speeded up" composition and insists on the fundamental role of "deep listening", which "leads one to altered states of consciousness. […] Anything that I play has to do with management, interpretation and decisions that are listening-based. So, what I'm hearing is then listened to, and that's what's guiding the playing". Oliveros seeks a "multi-dimensional form of listening" by continually asking herself: "How much can I listen to at once? How can I challenge myself as far as possible?" [21] The space in which the music takes place is just as important as the instruments, which are themselves spaces, since an instrument is a resonant sounding board. We have to listen not only to the other musicians we play with, but also to the space, which responds just as much as a musician.

4 Interview with Kosugi Takehisa

In November 2015, I met Kosugi Takehisa, and asked him to clarify the meaning of the term "dismantling the ego" [自我の解体] which appears in his book "Ongaku no Picnic" ([音楽のピクニック] Music Picnic) [22]. When he studied musicology between 1957 and 1961 at the National University of the Arts in Tokyo, Kosugi and was particularly interested in improvisation in jazz and so-called ethnic music. He was also looking for a method to integrate the sounds of the environment into music: "In the 1950s, we were looking for accidental collisions with sounds, for example daily sounds. At first, this 'everydayness' was not obvious in my music, but my colleagues were also interested in 'sound,' and we began using not only the sounds of music instruments but also environmental sounds [...] During the first concert of our 'Group Ongaku' [in September 1961], we played improvisation music, and also tape music. There is some kind of contradiction, but we used both concepts: on one side the ‘sound object’, on the other side 'improvisation'." [23]

At that time, the search for the "sound object" was necessarily accompanied by the practice of improvisation: the "concrete" sounds produced using "concrete" materials (magnets, wire or aluminum foil, etc) were used in improvised and experimental compositions, while searching for acoustic possibilities of the materials and devices. There was no fixed musical model at the time, and Kosugi had to look into other artistic practices: "From the end of the 1950's, the concepts of "Informel" and "futeikei" ([不定形] "without a fixed form", "indeterminate"), implying that the form of the artwork keeps changing, were widely used in visual and plastic arts. In music too: from a situation where the sounds and the self-consciousness were entirely fixed, it was then possible to move toward a place where the ego could be released more freely. I think these concepts have influenced many artistic areas. John Cage had also probably been thinking about how to liberate the ego from its boundaries. That's how I became interested in improvisation, and why I had to look at art and not only at music. […] Spontaneous music is somehow like calligraphy. […] Improvisation was our main concern: we wanted to produce spontaneous sounds, and our music was influenced by the changes in the movement of our own body or in the environment. [...] In fact, I was more interested in things that were constantly changing, than in something that was recorded and already finished. […] When you play, there is some kind of expressiveness, specific to that precise moment, which appears and sustains the music. It is always changing. I think that this aspect of improvisation is very important."

Kosugi then spoke about "manodharma" which has become a central idea in his work. Kosugi often used the term "Mano-dharma-electronic" for his electronic music works, after attending a carnatic music concert, and having read comments on the difference between "manodharma sangita" and "kalpita sangita". "When translated, 'manodharma sangita' becomes the 'Music of the Way of Thought'. 'Mano' is Thought. 'Dharma' is 'the Way' or 'the Law' (of nature). To me the ideal form of improvised music became the 'manodharma sangita'. Composed music is called 'kalpita sangita', where 'kalpita' has a connotation of 'imitation' or 'fake' [24] […] The idea of 'dismantling the ego', or of transcendency, is influenced by the specific concept of 'manodharma': usually, 'my' music is the music that 'I' play by myself. But 'manodharma' implies that the Ego exists as a cosmic existence beyond oneself, and a musician becomes a receiver that catches that cosmic existence. Improvisation reflects changes in time, or changes in season. Music comes out through the connections between the immediate environment and what transcends the ego. It is not 'my' music, but the music of a receiver that catches some presence in the universe, like a radio or a television. In short, a performance is something that catches the radio waves so that we can absorb them. That's what I mean by 'dismantling the ego'."

Kosugi has often insisted that improvisation was about reacting immediately to the situation in which we find ourselves. Like Pauline Oliveros who is constantly listening to her environment, Kosugi wants to constantly adapt to an environment that changes incessantly. "To catch the vibrations of the universe, and to adapt to the changing environment, is also something that holds me up. Improvising is a means to catch what is appearing. Therefore, by changing the environment, I can attract in myself changes in improvisation. In other words, when 'I' am performing, it is not only me who is performing. The performance is not only my own, because I am catching the environment. Because I am playing together with that environment, I am not just myself".

One can think with Kosugi that Dada's thought was also about "dismantling the ego". And Cage was also working in that direction, like Marcel Duchamp, Merce Cunningham, Nam June Paik, and David Tudor, among others.

5 Interpenetration of Different Temporalities

After the "structural improvisation" mentioned by John Cage, I would like to consider different approaches to musical time. In the 1990s, when computers became available, I created computer programs to compose and play with sounds that were repeated in ever-changing configurations. The set of sounds is present at the origin of the piece, but we cannot predict in what order and in what form each of them will be heard. The music does not move linearly, it is always renewed, if we listen to it as such, in other words if we want to be surprised. Such compositions lead nowhere and doubt any linearity, but they are at the same time combinable with other compositions and can integrate other time scales. They also have holes, or intervals, which invite to listen to the sounds of the environment.

A composition is thus a tool for listening to the world. The composer has made the sound listening device, and the listener is responsible for listening to the environment through this device. The composer is now in the same position with the performer and the listener. He defines a spaciotemporal framework, which settings can be modified when necessary, while remaining attentive to all the events that take place in an unpredictable way. It is therefore impossible to fix any score, as we cannot know what is going to happen. The compositions which I named "undirected" refuse to define a focus, or rather they pose that everything can be a focus, or a center. The composer-listener-interpreter directs his attention to each element without favoring any one, inside a situation that I call "interpenetration without obstruction of multiple and transparent spacetimes".

Can we suppose a "deeper" level of interpenetration? According to Wassily Kandinsky, "the same inner resonance can be achieved at the same time with different arts". [25] Artists wanting to get rid of boundaries, and aiming at the fusion between the arts, or between arts and sciences, have produced environments, installations, or performances, where separations disappear: their elements are complementary, interdependent, intermediate not only between media, spaces, or moments in time, but also between subjects and objects, between the subject and his environment.

Perhaps we could speak here of "pre-sensible": we cannot see a sound, but could we say that it is not totally invisible, that it is pre-visual? In the same way, would the music of a painting be pre-audible? There could be a communication between these differentiated registers, at the level of an undifferentiated ground [26]. Carl Gustav Jung has called this ground "synchronicity", which abolishes distances. Kimura Bin speaks of intersubjective relations between different subjects, and of intrasubjective relations between different levels of consciousness of a subject [27]. He calls "aida" [あいだ] the place where subjectivity is defined. According to Kimura, there is a common ground, and subjectivity is realized from its relationship to this common ground. Aida, or "space in-between", is the condition for the subject to be in relation with his environment, with the world, with the other. The other is both outside and inside the subject. Aida is also the condition for music to happen: musical time lies in the interval between sound and non-sound (silence). The true silence (primordial silence?) might be analogous to the Japanese koto [こと] (the intangible), whereas the objectified silence would be on the side of the mono [もの] (the tangible). Aida, the place of synchronicity, or simultaneity, would thus be situated between the mono and the koto, between the sound of silence and true silence. On this subject, John Cage spoke about reaching the "continuity of non-continuity" [28]. The different modalities of "improvisation" we have been considering might be useful to explore the "aida".

6 Oblivion

Having examined different approaches of music improvisation, we have noticed that for practical reasons there was often a necessity for structure in the conception and realization of improvised music, as suggested by Cage with his idea of "structural improvisation", or Wolff's instructions concerning the interactions between musicians. However, when Kosugi and Russell perform, they often don't need prior agreements about what is going to be performed, because such an agreement is short-circuited by their *discipline,* in the Cagian sense. An "ideal" combination of the various approaches described above might lead to realize a musical performance by instantaneously reacting to one's inner and outer environment and exploring the "aida" described by Kimura, while choosing from a vast corpus of musical vocabulary and structures, and at the same time being able to "reach the impossibility […] of transferring the memory imprint" [29] and forget all what has been learned and assimilated. This process of cognizance and oblivion might be the predisposition needed to focus on catching the vibrations of the universe.

References

1. English Oxford Living Dictionaries, https://en.oxforddictionaries.com, last accessed 2019/5/9
2. Cage, J.: Forerunners of Modern Music. In: *Silence*, Wesleyan University Press, 62 (1961)
3. Cage, J.: History of Experimental Music in the United States. In: *Silence*, 69
4. Cage, J.: ibid., 69
5. Stravinsky, I.: *Chroniques de ma vie*. Denoël-Gonthier, pp. 63--64 (1971)
6. Rosset, C.: *L'endroit du paradis*. Les Belles Lettres, collection Encre marine, 56 (2018)
7. Cage, J., in Lohner, H.: *Making of "One11" by John Cage* (1992), published as a DVD by Mode Records, 4'09"~ 4'35" (2006)
8. Cage, J., interviewed by Scheffer, F.: *How to get out of the cage*, published as a DVD by EuroArts Music International - Harmonia Mundi, 34'45"~ 36'35" (2012)
9. Cage, J. & Reynolds, R.: A Conversation. In: *The Musical Quarterly*, Vol.65, No. 4, Oxford University Press, 594 (1979)
10. Brooks, W.: Choice and Change in the Cage's Recent Music. In: *A John Cage Reader: In Celebration of His 70th Birthday*, by Gena, P., Brent, J., Gillespie, D. (eds.), CF Peters, New York, 95 (1982)
11. Dalhaus, C.: Composition et improvisation. In: *Essais sur la Nouvelle Musique*, Contrechamps, 194 (1972)
12. Cage, J., quoted by Shoemaker, B.: The Age of Cage. In: *down beat* (December 1984)
13. Zhuangzi, *The Inner Chapters*, http://www.indiana.edu/~p374/Zhuangzi.pdf, 26, last accessed 2019/5/9
14. Levi, J.: Les Leçons sur Tchouang-tseu et les Études sur Tchouang-tseu de Jean-François Billeter. In *Études chinoises*, vol. XXIII, pp. 416--417 (2004)
15. Yamaguchi K.: *Robot Avant-Garde*, Parco, pp. 77--84 (1985)
16. Cage, J. & Reynolds, R.: A Conversation. Op. Cit., pp. 580--582
17. Brooks, W.: Choice and Change in the Cage's Recent Music. Op. Cit., pp. 97--98
18 Cage, J.: The Future of Music. In: *Empty Words: Writings '73 -'78*, Wesleyan University Press, 183 (1979)
19. Wolff, C.: Conversation with Walter Zimmermann (1976). In: *Occasional Pieces: Writings and Interviews, 1952-2013*, Oxford University Press, 53 (2017)
20. Russell, J.: Somewhere There's Music, *Rubberneck 15*, (November 1993)
21. Oliveros, P., in Kalvos & Damian's New Music Bazaar, Show#52, 18 May 1996, http://econtact.ca/10_2/OliverosPa_KD.html, last accessed 2019/5/9
22. Kosugi T.: *Ongaku no Picnic*. Kaze no Bara, 60 (1991)
23. All quotes from this section come from an unpublished interview with Kosugi Takehisa, Ôsaka, November 26, 2015
24. Kosugi T.: *Ongaku no Picnic*. Op. Cit., 98

25.Kandinsky, V.: *Du spirituel dans l'art*. Denoël/Gonthier, 137 (1969)
26.See Dufrenne, M.: *L'œil et l'oreille*. L'Hexagone (1987)
27.See Kimura B.: *L'Entre – Une approche phénoménologique de la schizophrénie*. Millon (2006)
28.Charles, D.: *La fiction de la postmodernité selon l'esprit de la musique*. PUF, 26 (2001)
29.Duchamp, M.: *Duchamp du signe*. Flammarion, 24 (1994)

Improvisation: Thinking and Acting the World

Carmen Pardo Salgado
University of Girona
carme.pardo@udg.edu

Abstract. This article argues that musical improvisation practices represent a way of thinking and acting in the world that is diametrically opposed to that of a highly managed and automated society. Having accepted Houchard`s description of musical improvisation as the decantation of previously learned music, we will use Félix Guattari's eco-sophy to discuss improvisation practices as exercises that take place within a mental ecology and a social ecology. Within the mental ecology, one achieves the creation of an existential territory that needs to activate a state of oblivion to produce this decanting. Within the social ecology, this existential territory manifests itself in a collective temporality that can serve as a model to dismantle that other temporality created by an economic system that serves as the origin of individual and collective decisions.

Keywords: improvisation, mental ecology, social ecology, non-contemplative knowledge

1 Introduction

"You don't learn music – by which I mean non-preserved music; you don't make it 'work', it makes us 'work' – like language, chorea (including 'Saint Vitus' dance?), aesthetic gestures, etc. – – music is what grabs hold of us and returns us to our essential transformations, by subjecting and freeing our neotenic (un)condition which is that of the experimenter-inventor (not some musty keeper of inventories). [It returns us] to the other of our plural singularity, of our fluid and solidary solitude; to the immanence of our multiplicities by updating our idiolectal virtuality, freed from the ostracism of the ego, from its morbid shackles, from its stupid anguished turpitude, and from paralyzing mimesis.
This implies: not learning music, but de-chanting it - is that to say, decanting it?"[1]

[1] "On n'apprend pas la musique -j'entends celle de non-conservation ; on ne la 'travaille' pas davantage, c'est elle qui nous 'travaille' -comme la langue, la chorée (celle de 'Saint-Guy' y comprise ?), le/la geste plastique, etc. -, c'est elle qui nous happe et nous rend à nos vitaux devenirs, via l'assujettisement et la liberté de notre (in)condition néoténique, celle d'expérimenteur-inventeur (qui n'est pas inventorieur éventé), à l'autre de notre singularité plurielle, de notre fluide et solidaire solitude, à l'immanence de nos multiplicités actualisant notre virtualité idolectale, exonérée de l'ostracisme de l'ego, de son morbide carcan, de sa stupide turpitude angoissée, et de la paralysante mimesis.

Based on these words by the writer and improvisational musician, Jean-Louis Houchard, we will, in this article, consider improvisation as the practice of decanting music that has been previously learned. This process of decanting passes through two stages: first, within a mental ecology and second, within a social ecology.

This idea of improvisation as a process of decanting means we must distinguish between the way improvisation has occurred throughout the history of music, as for example in the Baroque period, and the way in which it has been presented, to a large degree, from the 1960s onwards. During the Baroque period, improvisation was carried out, particularly, during ornamentation, which consisted of the performer inserting notes interspersed between the main notes of the melody written by the composer. This practice became obsolete when composers, such as Johann Sebastian Bach towards the end of his life, started to write down the ornamentations they wanted and, consequently, to reduce the freedom given to the performer's improvisation. In this kind of improvisation, the decanting consists merely of a slight deviation from the melodic line proposed by the composer. The deviation produced from the heights of the sounds and/or the rhythms and durations, oscillates around the melody offered *a priori*, and remains always within the realm of what is considered music. In the 1960s, a quite different practice of improvisation had begun to take shape with, on the one hand, jazz experiments with improvisations that were not based exclusively on pre-defined harmonic matrices and, on the other, the open work and experiments with randomness in so-called contemporary music. At that time, among those involved in improvisation practices, composition was under question, as was the case with Franco Evangelisti of the *Gruppo de Improvisazioni Nuova Consonanza* (1964) in Italy, or Cornelius Cardew of the Scratch Orchestra (1969) in England.

"We don't need notes anymore" declared Anthony Braxton, while explaining that one important characteristic of improvisation was its ability to go beyond the note and reach "anotality".[2]

The interest in improvisation practices as a way of decanting music was reflected in the huge proliferation of groups that emerged, both in avant-garde jazz and in contemporary music. This decantation of what was musical, unlike the one previously mentioned with respect to the Baroque period, did not consist of a mere deviation, but also involved a filtering exercise in which the sounds are separated from the theoretical remainder that contains the architecture of the musical composition. The sounds could emerge outside the established theoretical structures.

In addition to the two groups mentioned above, the most outstanding in this respect were the *Musica Elettronica Viva* group in Italy (1966); the AMM (1965) and the London Musicians Collective in England (1975); the New Music Ensemble (1972), the Sonic Art Group (1966) and The Theatre of Eternal Music (1960s) in the USA; the Free Music Production in Germany (1968), and the *Taller de Música Mundana* in Spain (1978), among others.

Cela sous-entend : ne pas apprendre la musique, sinon pour dé-chanter -est-ce à dire pour décanter ?" Houchard [1], p. 25-26.

[2] Sparti [2]

It is not the purpose of this text to deal with the differences between improvised music in Europe and American free jazz, nor to discuss the different contributions each group has made, but rather, to note that this proliferation of groups in the 1960s, formed part of what Alvin Curran describes as "revolutionary and explosive artistic creativity", which also manifested itself in rock bands, electronic environments, hippies and potheads, both in Europe and the United States.

These revolutionary and explosive artistic practices are testimony to the exercise of decanting learned music. Alvin Curran himself explains it in an exemplary way in relation to his own practice and to the foundation of the group *Musica Elettronica Viva*:

"Shortly after that [*a year in Berlin*], hop in a car, go to Rome and that's where I've been ever since. That's where MEV, with Richard Teitlebaum, Fred Rzewski and myself got off the ground. Basically here we were very academically trained, academically directed, everyone said we were very promising composers, so presumably we would have been very promising composers, then we got to Rome and we figured something really wasn't quite right. And is just at the beginnings of the '68 revolution, about 1966. And for one reason or another there was enough curiosity, discontent and intuition among us to be able to form a group that rejected all forms of hierarchy, all forms of organisation, all forms of leadership. No director, no score, no knowledge of when the music might begin or end.

It was Tabula Rasa. It was erasing our whole background. Everything we were supposed to be, everything we were supposed to do and basically our whole cultural mission in life."[3]

This kind of decanting in improvisation is of particular interest to us here as it allows us to show, with examples, the stages we mentioned earlier in this decanting process: first, within a mental ecology and second, within a social ecology.[4] For the proposal presented in this article, we consider Guattari's ecological approach to be the most appropriate for a number of reasons. First, we can begin to understand improvisation practices in the context of the revolutionary and explosive creativity described by Curran, thus taking into account the political, ecological, artistic, technological or scientific issues, understood as a dynamic medium. Second, Guattari's ecosophy is not introduced as a theory, but as ecosophical practices oriented towards non-contemplative knowledge. Thirdly, ecosophy, by its own conception, implies decentralization in relation to ecology, politics, art, technology or science. Therefore, we consider that ecosophy, which embraces decentralized practices, constitutes the most suggestive model for discussing the process of decanting produced by improvisation.

Because of the complexity of these decentralization practices, in this text, this decanting can only be dealt with partially, but we think that it can provide a good model on which to base our future research work.

The analysis presented in this paper, focuses on improvisation as a permanent inquiry that affects the spiritual or mental and the political or social. Improvisation is

[3] England, P. [5]

[4] For more on Guattari's ecological approach, see: Guattari, [6]

conceived as non-contemplative knowledge because improvising is not knowing *a priori*; improvising is not foreseeing.

2 Improvisation and Mental Ecology

"I became interested [in improvisation] because I had not been interested. And the reason I had not been interested was because one just goes back to one's habits. But how can we find ways of improvising that *release* us from our habits?"[5]

This statement by American musician John Cage, helps us to understand one of the fundamental actions that occur when improvising.

Improvising means detaching oneself from acquired habits, although as we will see, this detachment is not complete.

In the 1970s, while John Cage was reading Henri David Thoreau, he reconsidered his distrust of improvisation. Through Thoreau's *Journal*, his book *Walden*, and the *Duty of Civil Disobedience*, Cage learned a great deal from this transcendental philosopher, who carried out his own *tabula rasa* regarding the sounds of civilization in order to immerse himself in the sounds of Walden. Through Thoreau, sounds, ecology, and politics, revealed themselves as part of the mechanisms that weave civilization together and, in turn, allow it to be questioned.

Around this time, Cage composed *Child of Tree* (1975), *Branches* (1976), *Inlets* (1977) and *Pools* (1977), respectively subtitled *Improvisation Ia*, *Improvisation Ib*, *Improvisation II* and *Improvisation IIa*. In the 1980s, he continued with *Improvisation III* (1980), *Improvisation IV* (1982), *Improvisation A*+B (1986) and *[C]omposed Improvisations* (1987-1990). In each of them, it is a matter of freeing oneself from habits. We will now deal with the first and the third works.

For the first, *Child of Tree*, named after a phrase in Joyce's Finnegan's Wake, the performer uses amplified plant materials. Cage indicates two of ten instruments to be used: a cactus to be played by plucking the spines with a toothpick or needle and a pod from a poinciana tree. Cage instructs the performer how, through random operations, to divide the eight minutes of the work into parts and how to divide the ten instruments between the different parts of the work. The performer improvises with the plant and the different instruments.

Inlets, for three performers, shells and the sound of fire, live or recorded, was premiered by the Merce Cunningham Dance Company in Seattle. The work has no temporal structure and was composed independently, in the usual mode of cooperation between Cage and Cunningham. Regarding the work, Cage only indicates that each shell has to be played only once. Using instruments that cannot be controlled, Cage finds a way to free himself from habits and to improvise.[6]

Through these works, Cage proposes processes open to the unknown and moves to a place where it is not possible to establish a duality between composition and

[5] Retallack, [7], p. 274.

[6] Regarding *Inlets* he states: "In the case of *Inlets*, you have no control whatsoever over the conch shell when it's filled with water. You tip it and you get a gurgle, sometimes; not always. So the rhythm belongs to the instruments, and not to you." John Cage in Cole, G., Caras, T. [8] pp. 76-77.

improvisation. This way of proceeding aims to produce the necessary insertion of art into nature. Cage was inspired by the thoughts of Ananda K. Coomaraswamy, who in his book, *The Transformation of Nature in Art*, affirms that the function of art is to imitate Nature in her manner of operation. According to Cage, nature is governed by processes in which chaos has a fundamental place. Nature's workings show the multiplicity of simultaneous processes that cannot be grouped together towards a common goal. Nature is not guided by a teleological process. Consequently, the imitation of nature will consist of imitating the way nature moves, the process.[7] Improvisation is understood as an exercise in experimentation and discovery of the unpredictability of nature and sounds. The works have nothing to do with the memory or the tastes of the composer, or that of the performers, they are simply searching. The work becomes concrete when improvisation takes place. As a result, there is no point in referring to a composer or an interpreter.

In reference to *Inlets*, Cage explains:

"What delights me in this thing… is that the performer, the improviser, and the listener too are discovering the nature of the structure … Improvisation … that is to say not thinking, not using chance operations, just letting the sound be, in the space, in order that the space can be differentiated from the next space which won't have that sound in it".[8]

To improvise is to not think, to not use the operations of chance and to let the sound be. To do this, it was necessary to make improvisation into a practice that frees memory and tastes. The ear can be directly linked to an instrument, without passing through a score or some other indication of the sound level. Furthermore, in these works the instruments are unknown.

So, improvisation means situating oneself in this realm of not knowing what is about to happen: to free oneself from habits. This liberation implies a decanting of the way in which one has learned to connect sounds with each other. In this sort of *tabula rasa* or emptying of memory and tastes, music appears upon a more open field of unsuspected possibilities. In this field, the musician discovers or invents sounds at the same time as he discovers and invents himself as a musician. This is an exercise in mental ecology that must first pass through this process of emptying oneself of knowledge about music and about oneself. This implies a kind of *putting-in-brackets*, a *leaving-it-in-suspense* to give oneself time to dismantle the forms in which, until now, this knowledge has been transmitted.[9] Secondly, the mental ecology of

[7] Cage, [9], p. 31; [10], p. 213.

[8] Cage and Reynolds [11], p. 581.

[9] An example of this was the improvisation of + - by Takehisa Kosugi and SCALE by Elena Asins performed by students from the Department of Art and Image Sciences of the Musashino University of Arts (Tokyo) and the Master in Sound Art of the Faculty of Fine Arts (Barcelona), in the Convent of Saint Augustine (Barcelona, 9/03/2019).
The choice of the pieces was the first step towards an approach to concepts and procedures that, despite taking place in globalized societies, continue to be alien to our culture. In this sense, and from the experience of the Barcelona students, previous work carried out on Takehisa Kosugi was important, but above all, the viewing, prior to the concert, of the film *Ma Space Time in the Garden of Ryoan-Ji* by Taka Iimura, with music by Kosugi – proposed by the musician Christophe Charles – was excellent preparation for that emptiness necessary

improvisational practices establishes an existential territory that is the territory of the improvisation process itself. This existential territory is, in Guattari's words, pre-objectal and pre-personal.[10] This means that this territory is not pre-constituted but is constructed with portions and aspects of what had previously been knowledge about music and about oneself as a musician. What once formed a chain of knowledge that followed the meaning of what was learned, now becomes fragments detached from the cause and purpose that had previously been established. Its objective is now, to activate itself in an unpredictable way in ephemeral sound organizations. In that earlier knowledge that previously formed the musical medium, other possibilities arise, other qualities that forge the fragile existential territory of improvisation. This territory is traversed by sounds and silences, by listening and by the gestures and decisions of the musician. To this, we could add other variables such as a possible audience, or the place where the improvisation takes place. As a result, as an appropriation of context that depends on a praxis, each improvisation is the gestation of a non-identical existential territory.

"The principle common to the three ecologies is this: each of the existential Territories with which they confront us is not given as an in-itself [*en-soi*], closed in on itself, but instead as a for-itself [*pour-soi*] that is precarious, finite, finitized, singular, singularized, capable of bifurcating into stratified and deathly repetitions or of opening up processually from a praxis that enables it to be made 'habitable' by a human project. It is this praxic opening-out which constitutes the essence of 'eco'-art." Félix Guattari, [6] p. 53

A mental ecology supposes, in consequence, the creation of an existential territory, of a house in which to live. In this house - of listening to sounds, silences and oneself - there are no universal rules. The essence of eco-art is the possibility of building houses, territories in which to live following construction lines different from those transmitted as the only valid ones. This dwelling does not prepare itself, to create habits. In the case of music, improvisation moves away from the boundaries of the musical field, which is considered a closed and strongly regulated arena. In the musician's example, he must create another commitment to the creation of a body that repeats and repeats in order to develop a muscular and mental memory that allows for the excellence that the delimitation of the musical arena has designated. To create an existential territory through the practices of improvisation implies, in consequence, dismantling the music field and the corporality that accompanies it.

3 Improvisation and Social Ecology

Improvisational practices are an example of what we call social ecology, which, following Guattari, consists of developing concrete practices which aim to modify

for learning how to listen to others. After this experience, the students of Barcelona created an improvisation group that still meets every week to think and play together.

[10] Félix Guattari [6], p. 54

and reinvent ways of being.[11] In the first place, what is being reinvented is the production of music as well as oneself as a musician. Secondly, improvisation is a re-composition of practices that is not limited to the field of music or the musician as autonomous entities, but also assimilates a way of understanding music in relation to politics, social, educational or other domains. This continuity between the musical, the social, the political or the educational is inspired, on the one hand, by the belief of a homology between musical organization and social organization. It is worth remembering, in this regard, the idea of tonal structure as being hierarchical and the twelve-tone musical technique as being a democratic organization.[12] On the other hand, the continuity between the musical, the social, the political or the educational, is approached from the belief that the practices experienced in the musical can serve as a model for the others. This ambivalence between homology and the consideration of musical activity as a model for the social, is particularly well-illustrated in many of the groups created since the 1960s, among which the practice of improvisation as a political act was included. In this way, one could highlight the explicitly Marxist line of the AMM group and the Scratch Orchestra to which Cardew belonged. Both were spaces for musical and political experimentation. Their aim was to destroy capitalism and the values imposed by the bourgeois class, such as the elitism of culture, or authoritarianism. The way to do this was to instill the space of the musical in the working class and, in consequence, banish it from the auditoriums and purge it of the bourgeois class.[13]

The example set by musicians has led to researchers proposing the use of improvisation as a model for changing social practices. For example, the research group behind the Adaptive Use Musical Instrument (AUMI) software has collaborated in an international project on Improvisation, Community, and Social Practice (ICASP), based at the University of Guelph, Canada, in which improvisation is being explored as a model for social change. This has led to improvisation practices being put forward as a counter-hegemonic model to the ideological system embodied by capitalism. The percussionist Edwin Prévost explains it succinctly:

"If we – as musicians and listeners – have any choice when confronting the morality of capitalism, then it must be to do rather than to be done to. We must decide who we are rather than be given an identity. In our freely improvised music there is the opportunity to apply a continual stream of examination. We search for sounds. We look for the meanings that become attached to sounds. And we have to decide – on the basis of observable responses – on the musical, cultural and social values that reside in whatever configurations emerge. The search is surely for self-invention and social-invention. This is an opportunity to make our world. If we do not act to make our world then somebody else will invent a world for us." Prévost [15] p. 58

[11] Guattari [6] p. 34

[12] Arnold Schoenberg compared tonality to a State in which the king is at the top of the summit and each chord is perceived as an actor whose aim is to take power: to become a tonic. Buch [12] p. 55. And for Adorno, as well, freedom in music and society are intimately linked. Adorno [13] p. 313.

[13] For more about this, see: Cardew [14]

According to Prévost, improvisation practices construct a world. The processes of invention in the musical are, at the same time, those of inventing a world of our own. To Prévost, the homology between the musical and the different levels of that world is possible thanks to practices that, starting from the sonorous, suppose a reinvention of oneself and of the social.

Creating this existential territory, which was highlighted in mental ecology, is now seen from another prism: that of social ecology. The collective character of a large part of improvised music is taken into account, as well as the fact that improvisation implies, as Prévost puts it, a continuous flow of examination. In relation to this flow and in order to shed light on some of the characteristics of the social ecology that underlies improvisational practices, we will address the temporal issue, but first let us examine the collective nature of improvisation.

In improvisational practices, the musicians are composers and performers, nullifying the usual distinction made in a musical field which, when it involves musicians playing collectively, must be under the authority of the score or conductor. In a collective improvisation, these distinctions disappear; the important aspects are the sound and the listening, which serve to establish a dialogical relationship that is not mediated by a score or a conductor. As explained by Jean-Luc Guionnet, during a collective improvisation, the sound produced has a dual role: to participate in the sonorous result, and to transmit information. The first information ensures the dialogical presence of the musician who emits the sound. This information is formed by gesture, and by attention, which accounts for whether it has been understood or not. The information "is the spreading influence". This means that they are in suspense and therefore far from the compositional activity understood in a traditional way.[14]

In this dialogical form the sounds are organized, but so are the musicians who collaborate in the improvisation. An attentive listening is necessary on the part of the members of the group in order not to be caught in situations of domination or authoritarianism. This involves a co-presence, in which each musician finds himself in that kind of 'emptying' prior to the creation of an existential territory that now contains the collectivity. Thus, improvisation can be conceived as a common project, while respecting the singularities of the participants. These singularities do not belong to the order of identity – fixed once – but rather to the ways of carrying out emptiness, individual skills and the ability to articulate knowledge and forgetfulness for the sake of producing unforeseen music.

Secondly, the revolutionary character of musical improvisation appears in the way one works with temporality. The practices of improvisation, unlike the performance of written music, are not intended to produce a closed work. Improvisation practices are processes:

"Written compositions are fired off into the future; even if never performed, the writing remains a point of reference. Improvisation is in the present; its effects may live on in the souls of the participants [...] but in its concrete form it is gone forever from the moment that it occurs, nor did it have any previous existence before the moment that it occurred, so neither is there any historical reference available." Cardew [14] p. 46

[14] Jean-Luc Guionnet [17] p. 133.

Improvisational practices do not correspond to a teleological time, which has been previously designed and which must be brought to an end. It is not an intentional time, but an opening of time as a duration, during which any sound can occur. In that time, memory and active oblivion are interwoven to create a music that eschews habits and dialogue, as is the case in a conversation. Unlike communication, conversation is a process that does not fit into any predetermined object. Conversing is turn-taking without any specific aim, accepting an unbalanced rhythm in order to show that one can journey through time – a time in which regularity is not the only possibility – meandering, or wandering, making life an indivisible time in which to go astray. This is the kind of conversation that takes place in improvisational practices. During that purposeless time in which listening and the production of sounds flow, the performers can also experience instants that produce something like a cut in that flow. At such instants, as Christine Esclapez explains, choice takes place. The choice to *stop* in order to continue the path in *another way*.[15] This choice is made while also taking into account the group being listened to. This is because, in these practices, the path is traveled at the same time as it is created; it is not fixed.

The existential territory that generates this kind of collective temporality differs from the existential territories prefigured by the forms of collective life that have been established according to the neoliberal system in which we are living. All too often, the patterns of family life, education, work, and leisure, follow a teleological tempo that guides the steps of the journey in a way that has been established previously. If, as Jonathan Crary explains, "the form of contemporary progress [is] the relentless capture and control of time and experience", then the time of collective improvisation would escape this control.[16] Evading these other collective times, prefigured by the dominant economic system, is what allowed improvisation practices to be seen as revolutionary and explosive actions. But this took place in the context of the 1960s. These days, the economic and ideological system has swallowed up all the previously revolutionary proposals. Using the slogans of the revolution, they can now advertise a car.

Companies such as Google, Whole Foods Market Inc., or Amazon, have begun to apply the principles of improvisation to their work groups and strategies. Numerous publications have reported on this phenomenon from the late 1990s to the present day. There are studies on the application of improvisation practices in theatre or music, of product development projects and how managers can create teams conducive to improvisation.[17]

Nowadays, a combination of planning and improvisation is seen as necessary; they appear to be the perfect allies for continued economic progress in the business world.

"For leaders of firms competing in dynamic and uncertain environments, it is important to recognize that solving unexpected problems and taking advantage of unexpected events is necessary to complement the traditional organizing principles of planning and anticipating with those of improvisation and responsiveness." Vera, Rodriguez-López [24] pp. 316-317

[15] Esclapez [20] p. 35.
[16] Crary [21] p. 40
[17] See: Conforto, Rebentish, Amaral, [22] pp. 8-10; Kamoche, Pina e Cunha, [23], pp. 733-764.

This way of working has attracted the attention of intellectuals who propose using improvisational practices as tools that help us to deal with – from another standpoint – the changes that our technologies are provoking in human behavior. A very interesting example is that of Bernard Stiegler, who has worked on questions of improvisation in relation to the processes of generalized automation arising from digitalization on a worldwide level. At the *Institut de Recherche et d'Innovation*, which he directs at the Georges Pompidou Centre in Paris, he initiated a project in 2014, which consisted of improvisational workshops that took place in different cities in Belgium and France. Musicians, artists, writers or philosophers, took part in these workshops, together with members of the public, in order to establish a dialogical relationship between critical engagement and improvisation. The outcome was the organization, together with jazz musician Bernard Lubat, of the fifth *Festival of the Unexpected Encounters* entitled *Think/Improvise in the city of Tournai* (Belgium) in August 2015. The theme of the Festival was improvisation in the context of a generalized automation that is turning individuals into automatons who, in the phenomenological sense of the term, have no protention; they do not anticipate; they do not decide.[18]

Unlike the approaches of the 1960s, where improvisational practices arose in opposition to those of the bourgeois economic system, Stiegler believes automation is always there and is therefore not to be rejected, but rather to be thought about; it is something one can use in order to wonder and improvise. Consequently, through reflection and improvisational practices, we need to question, in Stiegler's words, the relationships between the automatisms that inhabit us, both socially and mentally, and the possibility of de-automation. It is a matter of putting automation at the service of de-automation.[19] Stiegler [26] pp. 231-234.

It could be said, that the aim is to arrive at non-contemplative knowledge – not, however, by means of some previous knowledge that has become automated (such as the muscular memory generated after copious repetition, which eventually operates as if independent of the will of the musician); instead, what is needed is to understand the mechanism underpinning those automatisms in order to dismantle them; so that we can strive towards de-automation.

Repetitions are furnished by a system that establishes habits that inhibit consciousness. Improvisational practices, in themselves, constitute acts of liberation from those habits, but in this case, it is not only a question of decanting that which is musical, but also the magma that comprises the social.[20] Hence, it seems that improvisation practices can serve first as a means to understand something that is within us and of which we are unaware, i.e., automation, so that we can then proceed with the task of dismantling it. The most urgent undertaking of social ecology today is

[18] The term "protention" belongs to the phenomenology of Edmund Husserl, who uses it in relation to the consciousness of time in order to describe a mode of intentionality that looks forward to the future. Husserl [25].

[19] According to Stiegler, there is always improvisation. A musical interpretation of a Mozart work includes improvisation, and in the same way, there can be no improvisation without automation. For more on the Festival, see: http://penserimproviser.org/wp/

[20] Magma it was the Latin word for the "dregs" of an ointment (the sediment found at the bottom of the bottle after decanting the liquid. "Dregs" is what is left after decanting wine.

to liberate our social fabric and ourselves from the automaton that inhabits us. Improvisation is the first step.

References

1.Houchard, J.L., Humair, D.: Corps Flottants de l'Humeur Vitrée. Comprenant À la Verticale – À l'Horizontale de Mawen Noury. VOIX éditions, Elne (2018)
2. Sparti, D.: Dans le Signe du Son. Bruit, Voix, Corps et Improvisation. In: Images Re-vues Histoire, anthropologie et théorie de l'art, 7, Paysages sonores (2009). Available in URL: http://journals.openedition.org/imagesrevues/417
3. Lewis, G.: Too many notes: computers, complexity and culture in voyager. Leonardo Music Journal 10:33-39, (2000)
4. Curran, A.: Around the Avant-garde: Avant-garde and Counterculture. A Round Table with Terry Riley and Alvin Curran. Presented by Carmen Pardo. In: The Limits of Composition?, La Casa Encendida. Caja Madrid Obra Social, Madrid, pp. 73-94, (2008)
5. England, P.: Mass Ornaments. In: The Wire, November 2004 pp. 32-37. Avalaible: http://www.alvincurran.com/writings/wire%20unedited%20interview.pdf
6. Guattari, F.: The Three Ecologies. The Athlone Press, London and New Brunswick (2000)
7. Retallack, J.: Musicage: John Cage in Conversation with Joan Retallack. Wesleyan University Press, Hanover and London, (1996)
8. Cole, G., Caras, T.: Soundpieces: Interviews with American Composers, pp. 76-77. The Scarecrow Press, Metuchen, (1982)
9. Cage, J.: A Year from Monday. Wesleyan University Press, Middletown, Connecticut, (1968)
10. Cage, J.: M: Writings '67-'72, Londres, Calder and Boyars, (1973)
11. Cage, J., Reynolds, R.: A Conversation. In: The Musical Quarterly 65(4), pp.573-593, (1979)
12. Buch, E.: Métaphores Politiques dans le Traité d'Harmonie de Schoenberg. In: Mil neuf cent. Revue d'histoire intellectuelle, n° 21, pp. 55-76 (2003/1)
13. Adorno, Th.W.: Écrits Musicaux: Quasi una Fantasia. Gallimard, Paris, (1982)
14. Cardew, C.: Towards an Ethic of Improvisation. In: Treatise Handbook. Editions Peters, London, (1971)
15. Prévost, E. Free Improvisation in Music and Capitalism: Resisting Authority and the Cults of Scientism and Celebrity. In: Mattin, Iles, A.: Noise & Capitalism. Kritika, Donostia (2009)
16. Oliveros, P.: Adaptive use musical instruments. In: Deep Listening Institute, n.d. Web. 18 Feb. 2012. http://deeplistening.org/site/adaptiveuse
17. Guionnet, J-L.: Buttes-Témoins. In: Filigrane, nº 8, pp. 129-148. Éditions Delatour France, Le Vallier (2008)
18. Thomson, S.: The pedagogical imperative of musical improvisation. In: Critical Studies in Improvisation, 3(2). (2007)
19. Ward, C.: Anarchy in action. Freedom Press, London, (1982)
20. Esclapez, Ch.: Un Ange Passe… Cosmologie de l'Instant : des Êtres et de l'Univers. In: Esclapez, C. (dir.). Ontologies de la Création en Musique, vol. 2 Des Instants en Musique. L'Harmattan, Paris pp. 35-57, (2013)
21. Crary, J.: 24/7, Late Capitalism and the Ends of Sleep. Verso, New York, (2013)

22. Conforto, C., Rebentish, E., Amaral, D.: Learning the Art of Business Improvisation. MITSLoan Management Review, Vol. 57, No. 3, pp. 8-10, Reprint #57317 (2016) http://mitsmr.com/1SPz5DJ
23. Kamoche, K., Pina e Cunha, M.: Minimal Structures: From Jazz Improvisation to Product Innovation. Organization Studies, 22/5, pp. 733-764 (2001)
24. Vera, D., Rodriguez-López, A.: Leading Improvisation: Lessons from the American Revolution. Organizational Dynamics, Vol. 36, No. 3, pp. 303–319, (2007) doi:10.1016/j.orgdyn.2007.04.002
25. Husserl, E.: On the Phenomenology of the Consciousness of Internal Time (1893-1917). Dordrecht, Kluwer. (1991)
26. Stiegler, B.: Our Automated Lives: An Interview with Denis Podalydès. Liminalities: A Journal of Performance Studies, Vol. 14, No. 1 pp. 229-247, (2018)

Developing a Method for Identifying Improvisation Strategies in Jazz Duos.

Torbjörn Gulz[1], Andre Holzapfel[2], and Anders Friberg[2]

[1] Royal College of Music in Stockholm
[2] KTH Royal Institute of Technology
gulz@kth.se

Abstract. The primary purpose of this paper is to describe a method to investigate the communication process between musicians performing improvisation in jazz. This method was applied in a first case study. The paper contributes to jazz improvisation theory towards embracing more artistic expressions and choices made in real life musical situations. In jazz, applied improvisation theory usually consists of scale and harmony studies within quantized rhythmic patterns. The ensembles in the study were duos performed by the author at the piano and horn players (trumpet, alto saxophone, clarinet and trombone). Recording sessions involving the ensembles were conducted. The recording was transcribed using software and the produced score together with the audio recording was used when conducting in-depth interviews, to identify the horn player's underlying musical strategies. The strategies were coded according to previous research.

Keywords: *improvisation, jazz, improvisation strategies, musical interaction, musical communication*

1 Introduction

Improvising is an activity that raises great interest in many disciplines. What defines a free improvisation? When is prior knowledge required for an improvisation? The ethnomusicologist Bruno Nettl summarizes how the term improvisation has been treated historically in music research [1]. He describes how improvisation for a long time was ignored and how it later on, based on a Western music tradition, was used as an opposite activity to composition. The improvisation was to be considered as the musician's way of putting a personal touch to the music through interpretation. Regardless of how we approach the concept of improvisation, the aim for the definitions are different even if we limit ourselves to improvisation within the music field and exclude all other sorts of improvisation.

This study deals with jazz music where improvisation is very central, and therefore we need a definition of the term jazz improvisation. The British musicologist and guitar player Thomas Williams makes the following definition of improvisation in jazz that is useful in this study:

"Improvisation in jazz is an extemporaneous activity in which performers navigate complex mental processes to produce musical utterances that aim to be at once novel, spontaneous and interesting and also communicative (both to the wider ensemble and audience), well-structured and familiar." [2]

Within the jazz community, improvisation is based on the ability to contribute from a mutual platform within a mutual language. On top of this mutual language, which is partly described in traditional jazz theory, there are other elements that more refer to personal expressions of each musician.

There are different ways for jazz musicians to practice improvisation, from developing the more traditional form of storytelling [3][4], which is mostly built on a practice-based vocabulary of phrases, into the free jazz language with extended degrees of freedom. Analyzing the horizontal lines as melodies by using tools from melody analysis is one way to acquire some of the basic knowledge in jazz improvisation, and therefore there are many transcriptions of famous jazz musicians' solos available. In addition to books with transcriptions, there are also larger projects such as the Jazzomat project with a very extensive database of solo transcriptions [5]. This kind of analysis fits well with the jazz theory that is usually taught, a theory developed from extended classical music theory with the use of functional harmony.

A theme specifically addressed in this study are the strategies used by musicians when the first barriers in the form of limited instrument skills and insufficient knowledge of basic jazz theory have been overcome. With a starting point in Norgaard's [6] categorization of improvisation strategies, Williams [2] has developed five main categories of strategies on which this study is based:

1. Rhythmic (timing)
2. Pitch (chord scales, melody, harmony)
3. Timbral (i.e. instrument sound, effects)
4. Physical (physical klichs, phrases)
5. Dialogical (interaction)

The purpose of this study was mainly to design and test a methodology to uncover strategies in an unprepared jazz improvisation setting. The results from asking the participants were adapted to the coding system by Williams [2]. We will discuss the advantages and shortcomings of this coding system, with the goal to extend and apply it in an upcoming study with Swedish professional jazz musicians.

2 Experiment

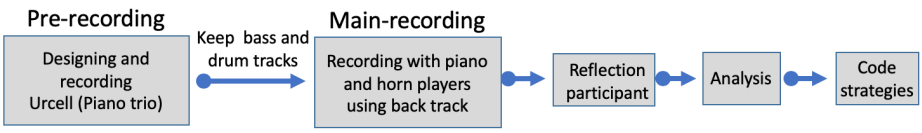
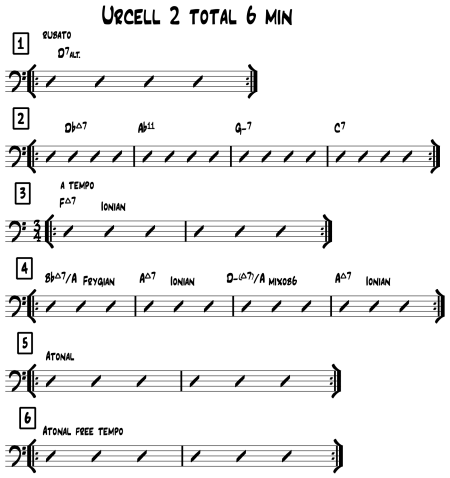


Fig. 1: Method applied in this study.

Figure 1 illustrates the process that underlies the study presented in this paper, which will be explained in the following parts of this section.

2.1 Music Material

In the preparatory stage five different piano trio schematic scores, called Urcells (Fig. 2), were performed and recorded, which later served as the musical platform for the duo improvisations (main-recording). The label Urcell is used through this paper to define the musical core of the experiment. The only instructions available in the score were implicit instructions, based on few chords or scales for the different six parts that lasted about one minute each. The five Urcells were deliberately composed with both easier and more difficult passages according to chord and chord scales. The studio recording with the piano trio was performed in a authentic recording environment by professional musicians. Every part of the Urcell was scheduled to be about one minute, with some variation due to the musical form. After the recording, the audio signals from the double bass were mixed down to a single track and all drums were mixed down to a stereo track, whereas the piano was muted. The six parts of the Urcells were:



1. Rubato, with one chord scale.
2. Rubato, harmonic movement with four chord scales.
3. A tempo, with one chord scale.
4. A tempo, harmonic movement with four chord scales.
5. A tempo, atonal
6. Rubato, atonal

Fig. 2: Example of the score from one of the Urcells used.

2.2 Main Recording

The main recording was performed with no visual contact between the musicians. The core of the music was the six-minute Urcell, based on the prerecorded material described in Section 2.1. The first four minutes of the main recording – before entering the Urcell – a freely improvised part without given conditions was played. This was merely for letting the musicians get into the music gradually. The Urcell together with the free introduction was to be seen as a whole composition-form. It was important, even though the recording was part of a

research study, to strive for a realistic artistic result. The recording environment resembled the situation within a real record session.

2.3 Participants

The participants in the study were four male advanced music students (age 22-26 years) from the jazz department of the Royal College of Music (KMH) in Stockholm, Sweden. The students were selected based on musical skills and their interest to contribute to the research study. All students were horn players and the four instruments used were trumpet, clarinet, alto saxophone and trombone. In addition to these students, the first author played the grand piano to complete the duo in the main recordings. The students faced a basically unprepared musical situation except for information about the aims of the experiment. The participants in the piano trio (pre-recording) were professional jazz musicians.

All participants filled in a consent form and were then informed about the basic idea of the experiment setup. The participants were also informed that a subsequent reflection would be asked for.

2.4 Equipment and Setup

The pre-recordings with the piano trio playing the Urcells were made in a studio at KMH. A Steinway grand, double bass and drum set were recorded on multitrack. All instruments were in separate rooms to avoid audio signals leakage. The main recordings of the study were conducted in a studio at the Royal Institute of Technology (KTH), using a Yamaha Disklavier with MIDI Output, two-channel USB sound card M-audio (fast track pro), and an Aston microphone (Spirit) for the horn players. The bass and drum tracks from the pre-recording were used as a back-track for horn and piano player in the main recording. The musician's reflection on the studio material and the analysis were recorded with a Zoom (H2n), while the analysis (score) was presented on a large monitor and the music was replayed in an audio system.

2.5 Data Analysis

Recording the piano part in the main recording via midi facilitated an accurate transcription of the piano using built in functions of Logic Pro X. The virtual sound of the grand piano (East-West piano platinum [7]) was found indiscernible from the acoustic sound by the participants. To obtain a simplified notation from the acoustic recordings of the horns, built-in features of Logic Pro X were used. The functions are inside the Flex Pitch module and are mostly used for pitch correction but in this case these functions were used as a transcription/visualization tool. Finally, the interpretation of the pitches was transformed into a MIDI-track and the score was presented.

This analysis was done immediately after the main recordings, and was accompanied by some transcription errors. Most of them were possible to adjust

in the flex pitch window by listening and transcribing traditionally by ear, and the obtained score was found to be sufficient as a starting point for reflection on the musicians strategies.

2.6 Reflection (Musician)

The reflection was performed in front of a monitor, and the principal focus was on the horn player's score. The score, along with the audio from the recording and with the ability to stop and rewind, provided sufficient information. Only limited guidance was provided for the reflection, in order not to influence the participant's reflection. Each reflection lasted for about 40 minutes. As a complement to the reflection, participants filled in a questionnaire with general questions regarding musical experience and background.

3 Results

The main results of the study concern method development. The method proposed here can, of course, be refined and further developed which is on the agenda for the continued research. Overall, the experimental situation was well adapted to describe some of the improvisation strategies in jazz, and specifically in treating how the improvisation output changes direction when conditions change. In this study the conditions differed due to the design of the Urcell.

With this study, the work also continues to categorize strategies within jazz improvisation. As role models, Martin Norgaard's [6] and Thomas Williams' [2] previous work are used to identify possible strategies and to group them. The main categories, (Rhythmic, Pitch, Dialogical, Timbral and Physical) are from the research of Williams. The students have been able to identify personal strategies by playing and then reflecting on their musical choices but they were not presented to the main categories by Williams in advance. Below examples[3] are provided of different strategies that have emerged during the study through the reflections by the participants.

a) Half step: the participant recognized that the note sounded out of scale and changed a half step up. In this case, the student discussed whether he actually perceived his note as the minor third, perceived a half note under the major third or just changed note because it sounded wrong, see Figure 3a.

b) Chromatic: the participant played chromatic melodic movements to postpone the resolution in the chord scale, see Figure 3b.

c) Pentatonic: the participant played a pentatonic scale. They are commonly used within the chord scale as well as outside the chord scale. This type of pentatonic scales belongs to what jazz musicians practice, attributing this strategy to the category Physical [2], see Figure 3c.

d) Tied notes: The participant kept to the same note when the chord changed and one strategy was to rapidly identify the function of the note. In this case

[3] Examples available at https://bit.ly/2YiNw9o

the participant immediately felt that the note (e♭) change from being the ♭9 in the D altered scale to the 9th in the new chord (Dbmaj7), see Figure 3d.
e) Stick to few notes: The participant decided to play few notes that belong to all chord scales in the example. In this example it is a trombone player. Fig. 3e.
f) Scale-wise: The participant quickly identified the scale and could in a more secure way play pre-produced patterns or new ones, see Figure 3f.
g) Copy notes: In this case the participant copied the note a from the piano and repeated it, see Figure 3g.

All strategies that emerged from the reflections were grouped according to Williams [2] as follows:

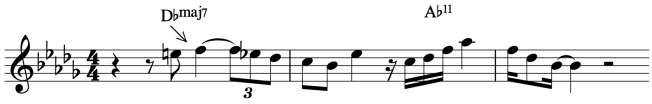
- Rhythmic: few notes
- Pitch: half step, scale-wise, chromatic, tied notes
- Dialogical: copy notes
- Timbral: idiomatic techniques
- Physical: phrases, pentatonic

There are advantages to grouping the strategies into main categories, and they were mainly applicable in this study. All categories are described by their names besides the category Physical, which is also a category developed partly from a guitar player's perspective. There is thus reason to review the categories and also to decide how the best representation of an individual strategy that falls under different categories would be.

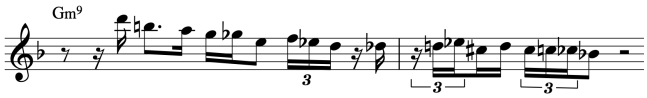
4 Discussion

The central theme of this pilot study has been methodological development since there are few previous studies to be based on. In an upcoming study with Swedish professional jazz musicians this method will be applied on a larger scale. In this kind of experiment it will always be essential to create a positive experimental situation, especially as this design of recordings can be experienced as demanding both in an artistic and a technical sense. Therefore, the introduction is particularly important, including soundcheck and rehearsal of some tunes. The surrounding environment can affect participants motivation for the experiment. It is often discussed among musicians how an optimal recording environment would look and work and several of the thoughts were directly transferable to this study. The students who participated in the pilot study were predominantly positive to the study. Above all, it was noticed how the method developed, in which the participants' performance was carefully analyzed and discussed, provided great opportunities for further development of the participants' musical development. The method can thus advantageously probably also be used as a teaching tool.

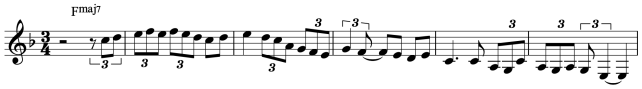
During the composition/construction process of the Urcell, it was essential that the whole piece obtains a functional musical form. Also, it should not be protracted to ensure that the concentration of the musician did not disappear. At the same time, for the data collection of the experiment, enough time must be



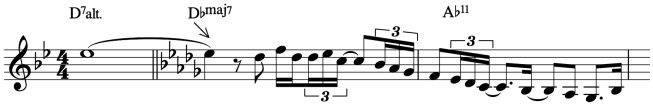
(a) Changing notes by a half step to arrive in the chord scale. (Clarinet score, non transposed)



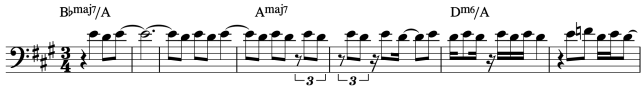
(b) Using chromatism to postpone the resolution. (Clarinet score, non transposed)



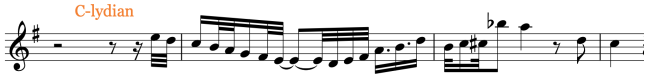
(c) Using pentatonic phrases to play inside or outside the chord scale. (Clarinet score, non transposed)



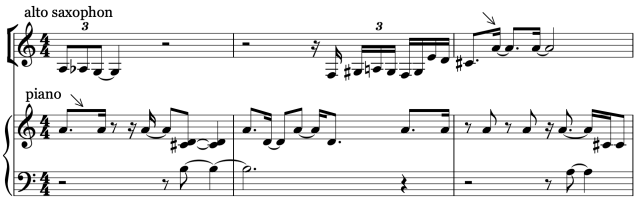
(d) How to understand the new function of the withheld note when the chord scale changes. (Clarinet score, non transposed)



(e) Example of sticking to few notes that are connecting chord scales and instead to use rhythmic variation. (Trombone score)



(f) Example of identifying the chord scale and to stay within it. (Trumpet score, non transposed)



(g) Note copying via dialogical interaction (alto saxophone and piano scores, non transposed)

Fig. 3: Examples of applied strategies.

available for each part. The structure with an Urcell consisting of six one-minute parts was found to be a reasonable length during the pilot study. The elements that are based on simple scales have in most cases been experienced as easy to decipher but have also provided an opportunity for a looser musical language where the musician takes off from a safe musical environment. The playing of parts that are moving faster from one chord scale to another create, with greater uncertainty, a more pending outcome. The accuracy of the analysis was high in this study compared to just recording and listening and the discussions were concrete and directly linked to the practice. One of the most interesting insights from the reflections and the survey is the impact that good ear training has on the participants self-confidence, which in turn leads to a larger palette of strategies.

5 Conclusion

To improvise in jazz is complex. This study shows the possibility to use a structured analysis of a musician's improvisation strategy through a method that provides a flexible framework but also retains space for the musician's freedom to create. The result offers a palette of strategies that move in several areas from music theory to individual choices. The main categories developed by Williams [2] seem well suited to this study, but will be further improved when conducting an extended study with more experienced jazz musicians where the frequency of occurrence of the various strategies will be investigated. The musical core of the recording, the Urcell, will be instrumental in these future experiments to bring to light the dependence of various strategies on how complex the parts are regarding rhythm and harmony.

References

1. B. Nettl, "Contemplating the Concept of Improvisation and Its History in Scholarship," *Music Theory Online*, vol. 19, no. 2, pp. 4–7, 2018.
2. T. Williams, M. Mermikides, and J. Barham, "Strategy in Contemporary Jazz Improvisation: Theory and Practice," Ph.D. dissertation, University of Surrey, 2017.
3. P. F. Berliner, *Thinking in jazz: the infinite art of improvisation.* University of Chicago Press, 1994.
4. S. Bjerstedt, "Storytelling in Jazz Improvisation: Implications of a Rich Intermedial Metaphor," Ph.D. dissertation, University of Lund, 2014.
5. M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart, *Inside the Jazzomat: New Perspectives for Jazz Research.* Schott Campus, 2017.
6. M. Norgaard, "Descriptions of Improvisational Thinking by Artist-Level Jazz Musicians," Ph.D. dissertation, University of Texas, 2008.
7. "http://www.soundsonline.com/pianos ." EASTWEST/QUANTUM LEAP.

Instruments and Sounds as Objects of Improvisation in Collective Computer Music Practice

Jerome Villeneuve[1], James Leonard[1], Olivier Tache[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France
[2] Independent researcher, Grenoble, France
jerome.villeneuve@gipsa-lab.fr

Abstract. This paper presents the authors' first attempt at a new (and unexpected) exercise: that of observing, contextualising and problematising their own collective Computer Music experiences. After two years practising emergent collective improvisation in private and public settings, which has led the authors to fundamentally reconsider both individual and collective musical creation, came the desire to methodologically deconstruct this process - one that they never anticipated and, until now, had never formalised. By starting from the very notions or performance and improvisation in the context of Computer Music, and crossing prolific literature on these topics with humble observations from their own experience, the authors then elaborate on what appears to them as the most enticing perspective of this creative context: the systematic improvisation of both their tools and sounds in an unique flow.

Keywords: Computer music, free improvisation, collective experimentation, instrument and sound improvisation, live patching

1 Introduction

This paper aims to question collective sound-improvisation in the context of digital technologies, based on a practice shared by the authors over the last two years. More specifically, from the perspective of three current or ex computer-music researchers all previously involved in various more traditional musical activities (from heavy metal to jazz), we will try to analyse how spontaneous musical interaction led us towards a free-form computer-based collective improvisation project named Orcæ - and the many interrogations that have emerged throughout collective practice and performance. As such, the paper is both a subjective testimony and a first attempt to methodologically deconstruct this shared practice in light of existing literature as well as the authors' musical and technological backgrounds. Starting from a brief overview and description of Orcæ's genesis and current creative process, we will work our way towards more fundamental questions such as: how do we behave collectively when improvising experimental electronic music? What can collective Computer Music performance *mean*? And can the notion of free-improvisation be extended to improvising the computer-based instrument itself?

2 A brief presentation of our study material: Orcæ

Orcæ is a trio of musicians composed of the authors that practices free collective music improvisation using mainly computers. Each of us has a different history of musical practices, including such diverse styles as heavy metal, jazz, reggae, *chanson française*, rock or electro-dub. Although we had never played music together before forming the band, we have a common experience as researchers in Computer Music and Digital Arts, having prepared PhDs and worked in the same team during a 5-to-10-year period. After several discussions regarding playing music together over the years, the project was initiated in January 2017.

2.1 Beginnings and gravitation towards free improvisation

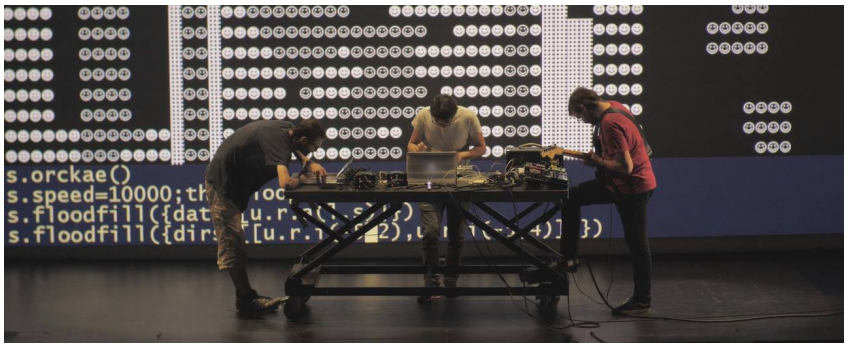
The initial purpose of Orcæ was to combine the authors' instrumental practices - namely guitar, keyboards and drums - with the idea of playing and performing post-rock music. Some songs were written beforehand, whereas other ideas were to emerge through recorded improvisation sessions, then to be transcribed and progressively fixed into song format. However, after recording and noting down a few improvised structures, attempts to reproduce them at a later time proved rather fruitless and frustrating: we all felt that something was "lost in translation", that re-exploring the same sounds was never as fun and exciting... Gradually, the electronic drumset became evermore drowned in post-processing and effects, before being abandoned in favour of a simple laptop. Similarly, fixed keyboard virtual instruments were replaced by a modular sound-synthesis environment, and the guitar became accompanied (and often replaced) by a laptop running sound transformation patches. Not only the music couldn't be written, but the instrument line-up itself was constantly evolving, sometimes expanding, other times shrinking. The progressive mutation was never planned, never completely grasped and never formally discussed by the players. After approximately 6 months of weekly sessions, our practice started to stabilise into the collective's present workflow.

The players each have different musical backgrounds, relationships towards improvised music (see Figure 1). It is worth noting that although we all come from a sound-synthesis technical background, most of our musical activity has been in current popular music genres (exception made of one or two electroacoustic fixed-piece compositions), and that only of us had any significant prior experience - or real interest - in free form (or *self-idiomatic*) improvised musical practice before this project. The music production tools used by each member within Orcæ also differ: Player A relies on *Max/MSP*[3], Player B creates mostly using *Reason*[4], and finally Player C uses *Ableton Live*[5].

[3] A modular patching environment for music and digital creation: cycling74.com/products/max

[4] The digital audio workstation (DAW) by Propellerhead: www.propellerheads.com/en/reason

[5] Arguably the most popular DAW for producing electronic music: www.ableton.com/en/live/



Player A	
Musical experience & training	Self-taught guitarist, formerly focused on heavy rock and metal: written music, rehearsed regularly and rendered "as is" live. Short spell in the Grenoble Conservatory's composition class.
Background	Software engineer & Computer Music PhD. Also sound engineer (mostly producing bands from punk to metal).
Link to improvisation	Small amount of jazz improvisation during first years of guitar playing - a skill now completely lost.
Instruments used in Orcæ	Electric guitar and various effect pedals Max/MSP patches with control surface.

Player B	
Musical experience & training	Self Self-taught musician, has successively played guitar, drums and keyboards in a now-defunct electro-rock band, before turning to solo electronic music production.
Background	Software engineer & software engineer, PhD in Computer Music, former Pure Data /DAW teacher.
Link to improvisation	Has practiced some free collective improvisation with his previous band (non-public jam sessions) and one-person improvisation as a way to compose electronic music.
Instruments used in Orcæ	Reason, and very recently Max/MSP.

Player C	
Musical experience & training	Formal education percussion, drums and piano, then jazz school. Drummer in various projects (ska-punk, big-band, raeggea, chanson française and Klezmer). Confidential electronic music composition.
Background	Computer Music PhD with a background in physics.
Link to improvisation	Systematic tendency to improvise when sitting behind drums, regardless of rehearsal or public contexts.
Instruments used in Orcæ	Ableton Live, always starting from the default patch at the begining of an Orcae session. Zero external controler.

Fig. 1. Above: Photo of a live performance in May 2018. Live-coded visuals were generated by Maxime Bouton and Emile Greis. Below: profile of each member of Orcæ.

2.2 Workflows and practices

Private Sessions start as soon as each member has connected their instrument to the main sound card and has a pair of headphones on. There are usually no directives exchanged between the members : we just start playing. One of us may occasionally propose a specific constraint (e.g. "let's not use any distortion today"), but most of the time such constraints are self-imposed as a way to avoid repetition and foster creativity. The session usually ends by an implicit common agreement, after anything from 40 minutes to well over an hour : sounds fade out, then one of us takes his headphones off, quickly followed by the others.

Public Sessions or performances were envisaged later (after nearly a year of playing together) and are handled a little differently. Before each performance, members usually exchange a few words about the global mood that the music may aim to achieve (although we rarely manage to stick to what we discuss beforehand). We are usually not aware of what other members have prepared (or in mind) for the performance, and enjoy having a few "tricks up our sleeves" for the others. Additionally, it is quite common for us to communicate verbally during public performances (e.g. "let's slow down") - while we hardly never do so in private sessions - particularly when trying to plan a "come down" for the closing minutes of the performance, as there are generally strict time limitations.

Multi-track Recording is systematic and has been since the very beginning of Orcæ, for both public and private sessions. This material is exploited to produce fixed audio tracks that we publish on the internet. The production process is kept as simple as possible so that the results resemble what can be heard live during a session, while filtering out certain inevitable moments were we are in more of a sonic research process than in a musical one. This work mostly consists in listening to raw material, selecting interesting portions and preparing them with limited editing and mixing as a stereo file (generally lasting from 3 to 15 minutes). We rarely desynchronise tracks, in order to keep the energetic cohesion from the collective improvisation. We are also rather attached to listening to the raw unedited recordings of our sessions, and have published a small amount of them, usually from public performances.

Collaborations have occurred regularly since the earliest stages of the project, through additional players occasionally performing with us as guests. We have worked with musicians and vocalists, video makers (in the context of producing spontaneous soundtracks for a short film playing in a loop during the session, or someone improvising live with us using a wide range of pre-recorded video capsules) and even live coders for real-time image generation. The latter have been a steady collaboration (both during private and public sessions).

3 Collective Computer Music Performance and Improvisation

In the following section, we will use Orcæ's creative process as a basis for analysing fundamental questions of performance and improvisation in collectively-practiced Computer Music. We propose to reflect upon these elements by combining various positions and results from the corresponding literature with interrogations and observations related to our personal practice. Although the acts of performance and improvisation are highly linked in this case, they will first be treated separately, as each bring forward a number of specific questions.

3.1 Performance

Performed Computer Music can designate any number of things. Our background lies in experimental music and academia. However we will consider here any public representation in which music is (at least seemingly) produced in the presence of a computer - englobing everything from electroacoustic contemporary music, to popular DJs, underground artists, to Laptop Orchestras and NIMEs[6].

Authenticity Computer Music performance in many of these contexts can spark a certain degree of confusion or skepticism among audiences since, as Andrew Schloss [1] remarks, it is not always possible for spectators to *"understand the performance from a direct/physical standpoint"*. It is indeed not trivial for an audience to know if all or part of the sounds that they are hearing are being generated through live performance, or if they are simply pre-recorded and then played back. To Schloss, this situation is deceitful: *"Tape music was boring to watch, but at least it was honest, with no false expectations of performance"*. He decries *"knob twiddling"* and other computer performance gestures that display no visual effort as things that should be either predetermined beforehand or discretely (and anonymously) performed behind the mixing desk.

Role of a Human Performer Schloss' primary focus is to bring back certain *theatrics* of effort and of corporeal causality from gesture to sound, a goal shared by much of the academic research on NIMEs, and by most of today's popular electronic music performers. One could argue that the question of ergonomics allowing the performer to finely control a Digital Musical Instrument is sometimes superseded by the question of representing and conveying "readable" gestural efforts for the sake of the observer/audience. One way or the other, designing meaningful corporeal links from gesture to sound in modern music is often problematic as a) one-to-one gesture-sound mappings are easily understandable but rarely sufficient for the musical discourse and b) complex gesture-sound mappings (e.g. triggering complex sound processes by means of relatively simple

[6] New Instruments for Musical Expression - conference: www.nime.org

gestures) can generate even more frustration from the observer, who is spectator to seemingly abstract gestures, perceptively unlinked to the sonic result.

For Guy Garnett [2], the human performer harbors more fundamental aesthetic consequences, such as the gestural nuance generally associated with human instrumental performance, rarely present in electroacoustic tape music:

it is more difficult to incorporate "performative" inflection into tape music, and therefore, for practical reasons, it becomes less likely to occur. [...] because [these subtleties] are difficult to produce, there is a definite tendency to avoid them.

Garnett also underlines the physical and cognitive constraints of human performance that affect the composer, the performer and the listener:

The performance gestures [...] must be cognizable: the performer must be able to get their mind around them in some way. The composer without physical limitations of performance can more easily convince himself or herself that they have created something real and comprehensible, whereas what they have may be an unhearable ideal. It is relatively easy to create algorithms that generate sounds whose qualities as music are inscrutable, beyond the cognitive or perceptive abilities of listeners.

One can therefore conclude that human performance in Computer Music is not only a question of adding readability to a restitution by expliciting (possibly caricatured) musical gestures, primarily directed towards an audience. Rather, human performance factors can be considered as fundamental structuring elements in the writing (or *thinking*, in the case of improvisation) of interactive Computer Music. As such, they are both meaningful and relevant even in the absence of performance, during any individual or collective creative processes.

Contexts & Expectations Considerations such as those presented above stem at least partially from heterogeneous conceptions of what could be identified as a *performance* according to composers, interprets or the audience - and, by extension, what each considers important or acceptable as a Computer Music performance. They certainly result in distinct expectations from each party towards the others. These co-expectations will tend to match if the context of the gathering is clearly specified: is it entertainment? A formal representation pertaining to a strongly-codified music genre? A scientific and technical proof of concept? An exploratory approach? An organic and open artistic journey? We are, of course, in no position to judge of the relative artistic validity of any of these contexts, however, finding which context Orcæ's performances "fit into" and which expectations we will confront has been a matter of trial and error.

Orcae's concerns Given that our public performances are constituted entirely of spontaneous real time improvisation, an inherent aspect of trust must be to installed between the audience and us. We invite them to embark on an open sound exploration, knowing fully well that it could be transcendental... or uncomfortable... or just very boring.

That being said, two recurring questions still obsess us and remain largely unanswered. The first, regularly expressed by the audience is : "who is doing what?". The subsequent second question being then "what should we explain to the audience beforehand, or what should we show, of what actually goes on during our performances?". Should we stick to a purely acousmatic listening experience and hide behind curtains, should we face the audience even though we barely seem to move during the whole session, should we visually project parts of our tools/screens (as a Causal augmentation) or should we go all out and build a complete dynamic scenography and audio-visual counterpart (as an Abstract augmentation)?

We seek for simplicity, and if we were to consider only ourselves (as is the case during private sessions - which in the end are simply performances in which we are both the performers and the listeners), we would not even think about anything but the sound for itself, disembodied of its producers. The fact that the performance aspect might not be seen at all or even known from the audience make little difference to us. But it clearly does for the audience. And while the literature largely states that fact, each one of our performances has been an occasion to measure it. We have played in various contexts and configurations (music only or working in collaboration with visual artists, playing on stage or amidst the listeners, fantastic to disastrous listening conditions, etc.) to various audiences, each time expliciting the bare minimum of our process (if we did so at all). Sometimes, the expectations of the audience converged with ours, some other time they did not. And the questions remain.

Further still, while the essence of our music may not have changed (too) drastically depending on these performance contexts, our subjective experience of each of them undeniably differs from the experience of private sessions. In other words, we don't feel any need to be considered as performers, however being put in a performing position/context significantly impacts our process.

3.2 Improvisation

In this section, we will not address the notion of improvisation in regard to the notion of composition. While the historical interest accorded to each has been very uneven (with a clear emphasis on composition, at least in western culture), numerous works have since proposed ways to formalize their relative positioning (see Sarath [3], Smith and Dean [4], Andy Hamilton [5]). We will restrain ourselves to the matter of musical improvisation involving computers. This specific field has seen distinct kinds of practices emerge and develop since the earliest ages of computer sciences. The first one would be to consider the *Computer-as-improviser*, able to generate structured musical information (e.g. MIDI then rendered by synthesizers). The second practice considers the *Computer-as-instrument* and emerges from the possibility of calculating real-time streams of synthesized or transformed audio data [6].

The practice of the authors within Orcæ is clearly positioned in resonance with the latter, in the sense that the computer is not perceived as an agent whose role is to respond creatively to the player's input (for instance by following

procedural rules), but is instead considered as an extensively controllable and reconfigurable instrument that allows for each parameter of each sound-producing process to be observable, editable or even stoppable at any given time[7].

Below, we will contextualize our approach and practice of computer music improvisation. From there, in the next section, we will posit that this context brings forth a second level of improvisation, referring to real-time design/deconstruction/re-construction of computer-based instruments.

Orcæ's improvisational process can be identified as pertaining to the codes of *self-idiomatic* music, as defined by Michael Bullock [7] (building upon Derek Bailey's term of *non-idiomatic* music):

> *self-idiomatic music is the concentration on sound-making actions for their own productive potential rather than in the service of representation of an external, received idiomatic identity.*

There is generally no prior agreement between players regarding any thematic, musical or stylistic directions, be it harmonically (no set key or preference for tonal or atonal material) or rhythmically (no shared tempo or clock synchronisation between machines). Sessions pass without any form of communication other than the sound itself.

Active listening is pivotal to collective improvisation[8] and may be even more so in this case, as each player's gestures are essentially limited to clicking, occasionally typing, and of course the infamous "knob twiddling". In other words, the sound is the only communication vector between players and the only means for developing a collective musical discourse[9]. As a result, the *who-is-doing-what* can occasionally become totally blurred, resulting in quite exhilarating moments in which each individual sound component dissolves into a greater entity and none of us are certain of the sound that we are each contributing.

Specificities of public improvisation Marcel Corbussin states that *"The possibility of failure is an intrinsic element of all improvised music"*, and while we certainly fail as much in private sessions as in public ones, the former feels much safer than the latter (at least for two of the three players). We tend to aim for a more "controlled" experience during public performances, often restraining

[7] This doesn't mean that we don't use emergent or chaotic sound processes (i.e. strongly nonlinear systems or feedback loops) but we don't consider the computer to be *improvising* in these cases - an electric guitarist controlling amplifier feedback is still a musician playing an instrument, even if the instrumental system is no longer passive in the mechanical/electrical sense.

[8] Marcel Cobussen [8]: "the constant process of decision-making that takes place during an improvisation is for a large part based on the listening attitude of the musicians involved."

[9] The degree of engagement and pleasure experienced during a public performance is then highly dependant on the quality of sound monitoring. Proper channels for this communication have to exist and low end systems can easily lead to frustration or even jeopardize the whole process.

our exploration of more "extreme" sonic territories, partly because there is a risk of producing uncomfortable sounds for the audience - but possibly because certain fears and inhibitions reappear in a public setting. Conversely, being in front of an audience yields a strong tension that develops focus and the feeling of flow, and as a result time seems to fly during public performances, to the point where it can be very hard for us to remember what actually happened[10].

Another consideration is that it may be difficult for a member of the audience to know, based solely on our performance, if the music is improvised or not - especially since we are not concerned with effort-based control gestures and so forth. Knowledge about how a piece of music was or is being produced has a significant impact on the listener's judgement [9], therefore we do ask ourselves if performances should start with a little disclaimer (*"be nice, it's impro!"*). Nevertheless, doing so may result in the audience focusing on us as performers, on what we are doing, how we are controlling sounds... whereas our aim is for the sound to be the object of interest in and for itself. As of yet we chose to say nothing beforehand.

Increasing risk - Alleviating failure One thing is for certain, for the audience as for ourselves: improvising computer music demands for perpetual richness, variety, curiosity and surprise. It seems that this posture must be considered on two different time-frames:

There is the time of the performance, during which we try to build an interesting exploration path for (with?) the audience. As expressed in section 3.1, the ability to match the expectations of an audience is of first common interest. This matter turns out to be even more crucial in the context of a free improvisation with computers. It leads Mazierska to express the following advice : "*[...] current electronic musicians are free to improvise, but if they want to keep their audience interested, they have to balance this need with the requirement to work with templates and observing traditions*" [10]. This statement brings us back to the inherent necessity of a (possibly unconscious) *common language* between performers, and between performers and audience. Nevertheless we find it important to emphasise on the widest possible interpretation of what these traditions or templates might refer to. We feel that they may include those from codified music, but also those closer to natural or evolving cultural hearing, such as our inherent tendency to relate to organic or artificial sounds through their potential to evoke the physical world, ambiances or even individuals.

And, there is - mostly for us, but maybe also for our most die-hard fans (if we have any) - a need to explore new creative fields on a wider time scale, from one collective public or private session to the next. This need was never defined as a prerequisite of our work together, it simply emerged from the fact that at some point, one of us would identify a routine coming from another (a recurring sound, effect, pattern, way to respond to or place himself in the macro form, etc).

[10] Ed Sarath [3]: *"The improviser experiences time in an inner-directed, or 'vertical' manner, where the present is heightened and the past and future are perceptually subordinated"*.

For some reason, being spotted was spontaneously felt as a personal failure in contributing to the collective effort of improvisation, and it progressively pushed each of us to rethink and reinvent our improvisation processes. This ultimately led to deconstruct the very notion of "musical instrument" and widen the scope of improvisation from sound only, to the low-level elements allowing us to produce it. In other words, one of our common practices now consists in starting from an entirely blank page/patch at the beginning of every session. As if the significant increase in risk was somehow the safest way not to fail our pairs or the audience.

4 Synchronous Improvisation of Instruments and Music

On the topic of the use of computers in improvised music, Frisk [11] expressed:

A computer does not have a sound but rather comprises the possibility of (nearly) any sound [...] to say that any sound is possible is not quite true [...] the kind of minute variation and dynamic change that constitute the very notion of a musical sound is still difficult to achieve on the computer. This is a programming challenge, a need to further develop synthesis techniques, but it is also a question of the interface between musician and computer[11]

Despite the ambiguous notion of *musical sound*, this statement relates directly to Orcæ's posture towards tools for Computer Music, and the need to investigate new paradigms of *improvising these tools*.

Indeed, since the late nineties research in software environments and programming languages for Computer Music have led to several tools - of both high and low level - that allow performers to program them and produce sound in real-time. The strongest movement that inherently carries such possibilities is Live Coding: *"Live coding is the writing of rules in a Turing complete language while they are followed, in order to improvise time based art such as music, video animation or dance"* [12]. It brings together a large community of performers/developers (for the most part academics or close to academy) around tools such as SuperCollider and Chuck. Another tool that worth mentioning is the Reactable [13], a - potentially collective - hardware interface that engraves physical objects with logical functions to be assembled on a visual display. In fact, it stands as a tangible version of visual programming environments such as PureData and Max/MSP, which allow for what can be called *live-patching* although it is not their most frequent use-case. And, finally another very interesting work relying on lower-level programming is the UrSound audio and multimedia engine [14].

Although the listing of these dedicated and often expert environments is relevant, luckily one does not need to graduate in computer sciences in order to explore this path. As a matter of fact, two of Orcæ's three members improvise using commercial software such as *Ableton Live* and *Reason*.

[11] This resonates directly with what Max Mathews stated at the dawn of Computer Music: the perspective of an infinite versus our ability to explore it in a sensible way.

4.1 Instrument Improvisation in Orcæ

Regarding musician and instrument in improvisation, Cobussen states :

> *The instrument does not simply yield passively to the desires of the musician. Likewise, he does not just bend it to his own will with no consideration to the resistance it offers. Rather musician and instrument meet, each drawing the other out of its native territory.*

Embracing the computer instrument as part of the improvisation process precisely offers a means to perpetually renew this play of resistance and exploration. In our case, novelty and personality in improvisation do not stem from virtuosity developed in relation to a given computer-instrument - something we see as difficult and possibly restrictive given the diversity and rapid evolution of available tools -, but in systematically "*(re)discovering the specific characteristics of their instrument, its unique and perhaps unexpected possibilities*".

In-session tool improvisation is mostly practiced by players B and C, as they generally start with blank *Live* or *Reason* workspaces. Although these environments possess advanced mapping possibilities for control surfaces, the players use almost exclusively the mouse & keyboard. Indeed, control mapping is usually used when performing with pre-structured musical environments, whereas player B and C's processes are qualitatively different in that they consist in creating work/creation environments in real time. To this day, the classic mouse/keyboard combination remains the most effective way to perform such operations.

Pre-session tool improvisation is practiced by Player A, whose main tool is Max/MSP. Live-patching entire instruments from scratch during sessions is rather tricky, so they are generally conceived beforehand. However, we still refer to them as improvised tools as they are often devised rapidly in the days or hours preceding a session, and are experimented in a *work-in-progress* state, tweaked, broken and fixed on the fly. The few of these experiments that stabilise over time into reusable tools are generally mapped to a control surface in order to facilitate exploration of the offered parameter space.

Imperfect digital tools In each of the above, one of the factors that drew us towards improvising Computer Music tools is the fascination for imperfections, a term often cited as a central aspect of musical improvisation [5]. So what if what we're doing is inducing hard audio-clipping? Or if we generate harsh digital artefacts every time we change the length of a delay line? In our short experience as recovering digital signal processing geeks / fresh young improvisers, all of these are simply spaces to explore and to work with - and while some choices may be frowned upon from a technical standpoint, who is to say that they can't be musically relevant?

5 Discussion

Although the positions advanced in this work inevitably fall into the domain of subjective evaluation and self-analysis of our own artistic process, it seems to us that the freely-improvised Computer Music context constitutes a unique and intriguing object of study. We believe that this improvisation scenario differs significantly from improvising on traditional instruments and that, in addition to the vast creative potential that it harbors, brings forth enticing interrogations as to multilayered improvisation paradigms and the creative exploration that occurs during musician-instrument interaction.

The format of this first attempt has led us to skim over a number of key considerations such as emergent collaborative creation, or multi-modal collective improvisation and performance. These will be for another occasion.

Finally, as a collective of improvisers who have *never* attempted to formalise to their creative process and approach to improvisation before writing this paper, one question remains... will this new awareness affect the way we improvise and perform together from this day on?

References

1. Schloss, W.: Using Contemporary Technology in Live Performance: The Dilemma of the Performer. Journal of New Music Research, Issue 32, p239-242. (2003)
2. Garnett, G.E.: The Aesthetics of Interactive Computer Music. Computer Music Journal, Vol. 25, No. 1, Aesthetics in Computer Music, pp. 21-33 (2001)
3. Sarath, E.: A New Look at Improvisation. Journal of Music Theory, Vol. 40, No. 1 (Spring, 1996), p1-38 (1996)
4. Smith, H., Dean, R.T.: Improvisation, Hypermedia and the Arts since 1945 (Performing Art Study). Amsterdam: Harwood Academic Publishers (1997)
5. Hamilton, A.: The art of improvisation and the aesthetics of imperfection. The British Journal of Aesthetics, Volume 40, Issue 1 (2000)
6. Dean, R.T.: Envisaging improvisation in future computer music. In R. T. Dean, ed. 2009. The Oxford Handbook of Computer Music. New York: Oxford Univeristy Press, p133-147 (2009)
7. Bullock, M.T.: Self-Idiomatic Music: An Introduction. Leonardo Music Journal, Volume 43, p141-144 (2010)
8. Cobussen, M.A.: Improvisation. An Annotated Inventory. New Sound, Issue 32, p9-22 (2008)
9. Canonne, C.: Listening to Improvisation. Eastern Division Meeting of the American Society for Aesthetics, Philadelphia (2017)
10. Mazierska, E.: Improvisation in Electronic Music—The Case of Vienna Electronica. Open Cultural Studies. 2. 553-561 (2018)
11. Frisk, H.: Improvisation, Computers, and Primary Process: Why improvise with computers?. New Sound, Issue 32, p107-118 (2008)
12. McLean, A., Wiggins, G.: Live Coding towards Computational Creativity. In Proceedings of 2010 Conference, Lisbon (2010)
13. Kaltenbrunner, M., Geiger G., Jordà S.: Dynamic patches for live musical performance. In Proceedings of 2004 NIME Conference, Singapore (2004)
14. Essel, G.: UrSound, Live Patching of Audio and Multimedia Using a Multi-Rate Normed Single-Stream Data-Flow Engine. ICMC 2010 Proceedings, Lisbon (2010)

Melody Slot Machine: Controlling the Performance of a Holographic Performer

Masatoshi Hamanaka[1]
[1] RIKEN
masatoshi.hamanaka@riken.jp

Abstract. This paper describes our interactive music system called the "Melody Slot Machine," which enables controlling the performance of a holographic performer. Although many interactive music systems have been proposed, manipulating performances in real time is difficult for musical novices because melody manipulation requires expert knowledge. Therefore, we developed the Melody Slot Machine to provide an experience of manipulating melodies by enabling users to freely switch two original melodies and morphing melodies.

Keywords: A generative theory of tonal music, time-span tree, melody morphing method, holographic performer

1 Introduction

This paper describes an interactive music system called the "Melody Slot Machine," which enables musical novices to manipulate melodies to be played by performers on a holographic display. In the Melody Slot Machine, melody segments are displayed on a dial, and the melody to be played can be switched by rotating the dial. The variations of melody segments are composed on the basis of the generative theory of tonal music (GTTM) [1], and thus, switching the melody segments keeps the overall structure of the melody and only changes ornamentation.

We believe that a computer that understands music deeply must be developed to achieve a composition support system that enables various music operations. Therefore, for 15 years, we have been implementing the GTTM, proposed by Fred Lerdahl and Ray Jackendoff in 1983 [2–5]. The performance of GTTM analysis by computers has been dramatically improved by introducing deep learning [5].

The time-span tree, which is the deep structure obtained as a result of GTTM analysis, shows the relationship between the main notes and the ornamentation notes structurally. The main advantage of using the time-span tree is that it is possible to reduce the notes of a melody by decreasing the ornamentation notes step by step while maintaining the main structure of the melody. One example of a melody operation by using the time-span tree is melody morphing, which generates an intermediate melody from two input melodies. Previously, we proposed a melody morphing method for generating an intermediate melody of two melodies as a melody operation that uses the time-span tree [6, 7]. We are also studying an operation called flip to invert the time-span tree [8]. We are planning to achieve the operation of various melodies in the future by using time-span trees.

We believe that melody manipulation by the time-span trees can not only enable musical novices to manipulate melodies but also improve the composition efficiency of professional composers. To assess whether the former is feasible, we constructed a system called the Melody Slot Machine.

The paper is organized as follows. Section 2 describes related work, and Section 3 gives an overview of the Melody Slot Machine. Section 4 introduces the time-span tree of GTTM, and Section 5 explains the melody morphing method that we proposed. Section 6 explains how we generate the melody for the Melody Slot Machine, and Section 7 explains the hardware implementation of the Melody Slot Machine. Section 8 describes an experiment on its brief implementation, and Section 9 concludes with a summary and an overview of future work.

2 Related Work

Our aim in building the Melody Slot Machine is that the user freely controls the holographic virtual musician. Although many interactive music systems have been proposed, we only discuss virtual music players. The Band-out-of-the-Box enables customized interaction between a live, improvising musician and the computer [9]. The Virtual Musical Multi-Agent System [10] and OMax Brothers [11] enable improvising with a human performer on the basis of a multi-agent architecture. Continuator makes it possible to search for melodies that are an accurate continuation of a melody played by human performers by augmented Markov models [12]. The Guitarist simulator enables a jam session with a virtual performer that statistically learns the reaction model from a real human player. However, the outputs of these systems are only performance sounds, and the performers cannot be seen [13].

On the other hand, the VirJa Session enables visualization of performers [14]. However, the visualizations are synthesized by computer graphics, which can only show simple motions such as foot-tapping or body-rocking. The Melody Slot Machine uses footage of performers actually taken by video camera.

As for systems using time-span trees, melody summarization systems [15], performance rendering systems [16], and melody prediction systems [17] have been proposed so far. Among them, a real-time interactive system is a melody prediction system. The melody prediction system assists novices with musical improvisation by displaying the predicted notes on the piano lid. When the novice finds it difficult to continue playing the melody, she/he can continue the improvisation by playing a note displayed on the lid, without impairing tonality.

On the one hand, the melody prediction system cannot be used by everyone because it requires a piano keyboard to be played. On the other hand, by using a melody operation method that involves rotating a dial, the Melody Slot Machine can be used by everyone.

The musical dice game combines the melodies selected by dice in order [18]. The Melody Slot Machine is inspired by the dice game in terms of changing the melody. However, the melody composition method is very different between the two. Melodies in the dice game are designed very elaborately so that they sound natural when combined in any order. On the other hand, our melody morphing uses the results of music structural analysis by GTTM. Even if the melody is changed, the essential part of the structure is the same, and only ornamentation notes are changed.

3 Melody Slot Machine

The Melody Slot Machine, an interactive music system that enables controlling performers on a holographic display, has three features.

3.1 User Friendly Interfaces

To enable anyone to easily manipulate melodies, we used a dial type interface that makes it possible to replace a part of the melody segment (Fig. 1 and 2). There is a rectangular hole in a part of the acrylic board sandwiching the score, and the dial interface on a tablet can be operated with fingers through the hole. When the red lever on the right side of the score is pulled down, all the dials rotate, and one of the melody segments on the dial is randomly selected.

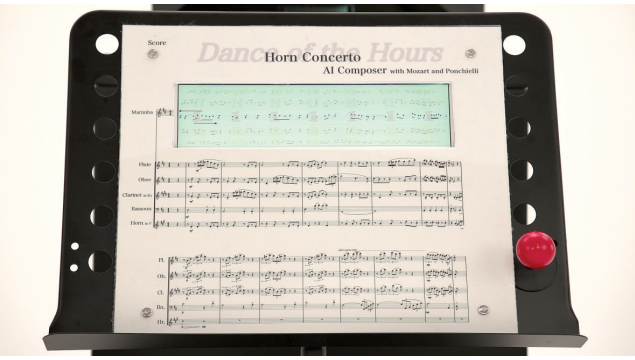


Fig. 1. Slot dial and lever

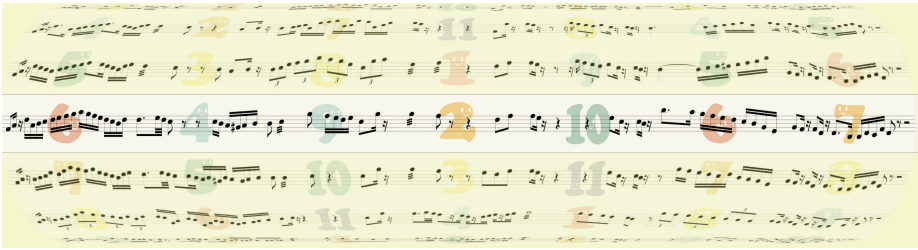


Fig. 2. Slot dials

3.2 Easy-to-Understand Operation Results

We prepared a display showing a performer so that the result of the operation can be confirmed visually as well as aurally. A holographic display was used to show the performer so as to increase the feeling of presence (Fig. 3). The users can feel like they are controlling a performer by operating a melody.

3.3 Improving the Feeling of Presence

We recorded all the performance sounds so as to increase the feeling of presence. For the recording, we used a studio with very little reverberation; otherwise, only the reverberation of the preceding sound enters the beginning of the melody segment because of the melody splitting into segments. Reverberation is added when the melody is played. Three pairs of speakers were installed, and panpot and reverb were set for each direction so that the hologram seemed to be a real performer (Fig. 4). When putting one's head between two pairs of speakers, both the sound and video enhance the feeling of presence.

Fig. 3. Holographic display

Fig. 4. Three pairs of speakers

4 Time-Span Tree of GTTM

Melody morphing uses time-span trees obtained from the results of GTTM analysis. The GTTM consists of four modules, each of which assigns a separate structural description to a listener's understanding of a piece of music. As shown in Fig. 5, the four modules output a grouping structure, metrical structure, time-span tree, and prolongational tree. The time-span tree is a binary tree, which is a hierarchical structure representing the relative structural importance of notes that differentiate the essential parts of a melody from ornamentation.

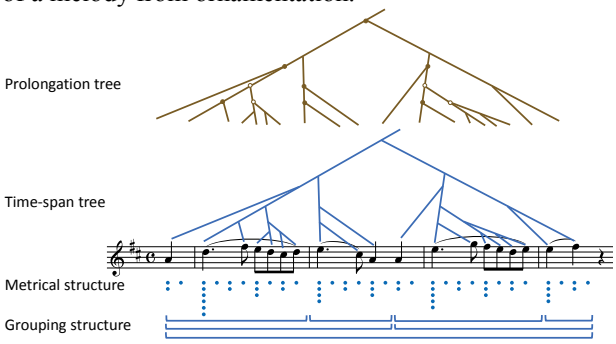


Fig. 5. Prolongational tree, time-span tree, metrical structure, and grouping structure obtained from GTTM analysis

4.1 Abstraction of Melody

Figure 6 shows an example of abstracting a melody by using a time-span tree. The figure includes a time-span tree from melody D, which embodies the results of GTTM analyses. In the time-span tree, important notes are connected to branches nearer the root of the tree, whereas unimportant notes are connected to leaves. We can obtain an abstracted melody, E, by slicing the tree in the middle (line E) and then omitting notes whose branch connections are below line E. In the same manner, if we slice the tree higher up at line F, we can obtain an even more abstracted melody, F. We can regard this abstraction of melody as a kind of melody morphing because melody E is an intermediate melody between melodies D and F.

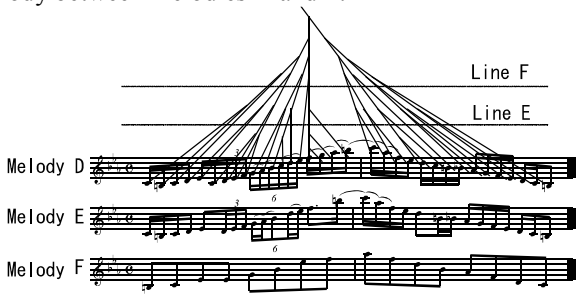


Fig. 6. Abstraction of melody

4.2 Primitive Operations of Time-Span Trees

To implement melody morphing, we use the primitive operations: the subsumption relation (written as ⊑), meet (written as ⊓), and join (written as ⊔) [19]. As shown in Figure 7a, subsumption represents the relation by which "an instantiated object subsumes an abstract object." For example, the relationship among TD, TE, and TF, which are the time-span trees (or reduced time-span trees) of melodies D, E, and F in Figure 6, can be represented as follows:

$$TF \sqsubseteq TE \sqsubseteq TD \quad (1)$$

Figure 7b illustrates the meet operator, which extracts the largest common part or most common information of the time-span trees of two melodies in a top-down manner. Finally, Figure 7c illustrates the join operator, which joins two time-span trees in a top-down manner as long as their structures are consistent.

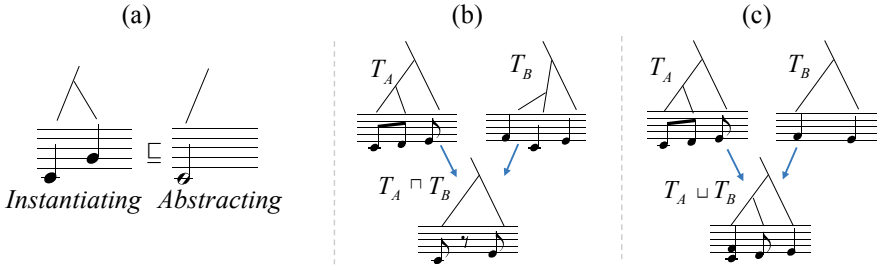


Fig. 7. Examples of subsumption ⊑, meet ⊓, and join ⊔ operations

5 Melody Morphing Method using Time-Span Tree

In this section, we explain the GTTM-based melody morphing method we proposed in 2008 [6, 7]. The initial melody A, target nuance melody B, morphing result melody C, and melody morphing method must meet the following conditions: 1 and 2 for melody C, and 3 and 4 for the method.

1. A must be more similar to C than to B, and B must also be more similar to C than to A.
2. When B is the same as A, C will be the same as A.
3. The output of multiple Cs depends on parameters that determine the levels of influence of the features of A and B.
4. C will exhibit monophony if A and B are monophonic.

5.1 Overview of Melody Morphing Method

Morphing means to change one image into another through a seamless transition. For example, a morphing method for a facial image can create intermediate images through the following operations.

1. Link characteristic points such as the eyes and nose in two images, as shown in Figure 8a.
2. Rate the intensities of shape (position), color, and so forth in each image.
3. Combine the images.

Similarly, our melody morphing method creates intermediate melodies through the following operations.

1. Link the most common information of the time-span trees of two melodies, as shown in Figure 8b.
2. Abstract the notes of a melody in a differing branch of the time-span tree by using the melody divisional reduction step of our melody morphing method.
3. Combine both melodies.

Figure 9 illustrates the melody morphing method.

Fig. 8. Examples of linking two images and two melodies

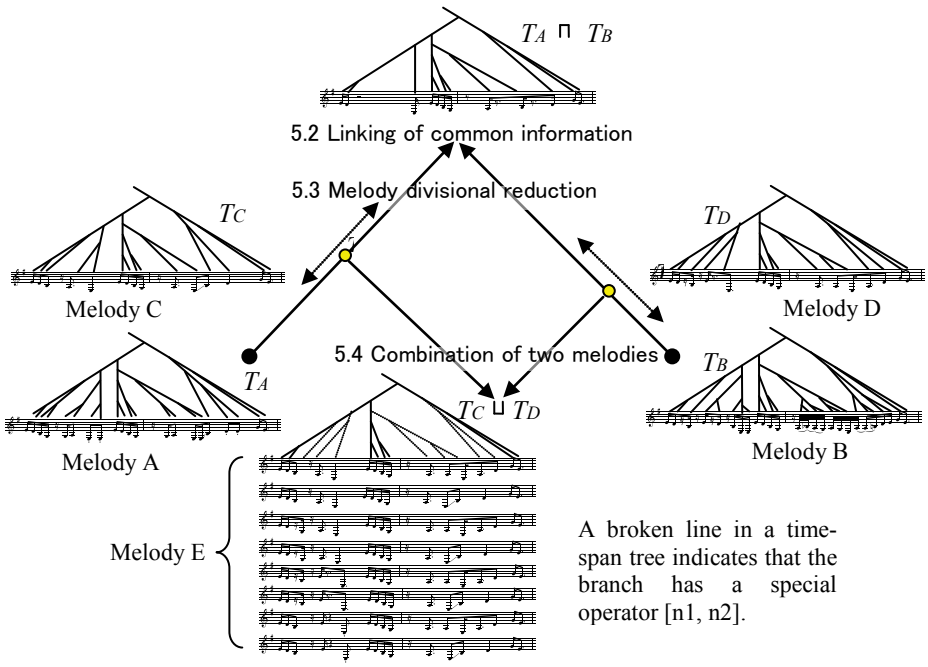


Fig. 9. Overview of melody morphing method

5.2 Linking of Common Information

By using the respective time-span trees T_A and T_B from melodies A and B, we can calculate the most common information $T_A \sqcap T_B$, which includes the essential parts of not only A but also B. The meet operation $T_A \sqcap T_B$ abstracts common notes from T_A and T_B, and the discarded notes are then regarded as the difference information of T_A and T_B.

When calculating $T_A \sqcap T_B$ by extracting the largest common part of T_A and T_B in a top-down manner, the result may change depending on whether octave notes such as C4 and C3 can be distinguished. If we discriminate octave notes, then C4 ⊓C3 will be empty, denoted as $\perp$. On the other hand, if we do not discriminate octave notes, the result is just C, which abstracts the octave information. Here, we regard a note and its octave as different notes, because processing is difficult if the octave information is not defined.

5.3 Melody Divisional Reduction

We next consider that the difference information of T_A and T_B includes features not present in the other respective melody. Therefore, we need a method for smoothly inserting or removing such features. The melody divisional reduction step of our melody morphing method abstracts the notes of the melody in the difference information of T_A and T_B by applying the abstraction described in Section 3.1.

Using this method, we can acquire melodies Cm (m=1,2,…,n) from T_A and $T_A \sqcap T_B$ with the following algorithm. The subscript m indicates the number of notes in the difference information of the time-span trees that are included in T_{Cm} but not in $T_A \sqcap T_B$.

Step 1: Determine the level of abstraction.

The user selects a parameter L that determines the level of abstraction of the melody. L can range from 1 to the number of notes in the difference information of the time-span trees that are included in T_A but not in $T_A \sqcap T_B$.

Step 2: Abstract notes in the difference information.

The note with the fewest dots in the difference information is selected and abstracted. The numbers of dots can be acquired from the GTTM analysis results [1–5]. If two or more notes share the fewest dots, we select the first one by reading the music left to right.

Step 3: Iterate.

Step 2 is iterated L times.

Subsumption relations hold as follows for the time-span trees T_{Cm} constructed with the above algorithm:

$$T_A \sqcap T_B \subsetneq T_{Cn} \subsetneq T_{Cn-1} \subsetneq \ldots \subsetneq T_{C2} \subsetneq T_{C1} \subsetneq T_A. \qquad (2)$$

In Fig. 9, nine notes are included in T_A but not in $T_A \sqcap T_B$. Therefore, the value of n is 8, and we can obtain eight intermediate melodies Cm (m=1,2,…,n) between T_A and

$T_A \sqcap T_B$. Hence, melody *Cm* attenuates features that occur only in melody A but not in B. Figure 10 illustrates this process.

In the same way, we can obtain melody D from T_B and $T_A \sqcap T_B$ in the following manner:

$$T_A \sqcap T_B \subsetneq T_D \subsetneq T_B. \quad (3)$$

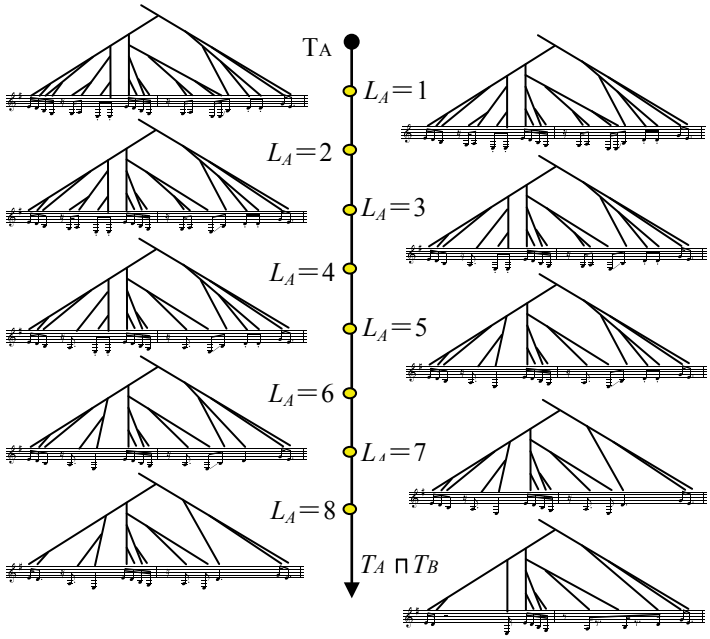


Fig. 10. Examples of subsumption ⊑, meet ⊓, and join ⊔ operations

6 Implementation

The holographic display (Dremoc HD3) can be viewed from three directions by using three glass panes with semitransparent film and reflecting the display installed on the top of the device. Figure 4 shows the hardware implementation of the Melody Slot Machine.

Speakers were installed so as to sandwich the proper head position in all three directions to urge the user to view the holographic display at the best position possible. We used Anthony Gallo Micro Satellite speakers, which have high-sound quality even at low volume. The sound is converted to three stereo SPDIF signals using the audio interface (RME Digiface USB) connected to the computer and then connected to three amplifiers.

The system is controlled using the max/MSP on the computer. The sound files are recorded in advance for each track and are divided into segments, and saved files are used. The length of each segment is the closest to one bar out of the grouping

structures of GTTM, which are matched from two input melodies of morphing. When the user sets the dial numbers on the tablet, the files are played in the order concatenated on the max/MSP. At the joints of the sound file, the volume is cross faded with a width of 30 milliseconds so that the volume level changes smoothly.

The accompaniment track is a separate file and is played back in a loop. Mastering parameters are set for each of the three directions by a sound designer using a plug-in installed on the computer so that it feels as if the sound is heard from the performer of the holographic display.

The video signal is connected to the holographic display with an HDMI cable via the graphics card. The video data have been taken in advance from three directions for all melody tracks. When the user sets the dial numbers on the tablet, the system plays in accordance with those numbers. The video is large in size compared to that of the audio, and it takes more time to start playing the file, making it more complicated than sound processing.

First, all the video files are concatenated into one file, and two copies of that file are written to a disk as file 1 and file 2 (Fig. 12). Then, during the playback of video A in file 1, if the next video to be played back is B, file 2 seeks the playback position of video B, and playback occurs immediately. At the moment video A ends, it releases the connection of the renderer connected to file 1, connects to file 2, and plays video B. The switching of this renderer ends within one frame, which is within 33.3 milliseconds, so a smooth connection without dropped frames is apparent.

We implemented this algorithm by using the VID-DULL video engine in MAX/MSP and Apple ProRes 422 video codec with a frame rate of 30 fps.

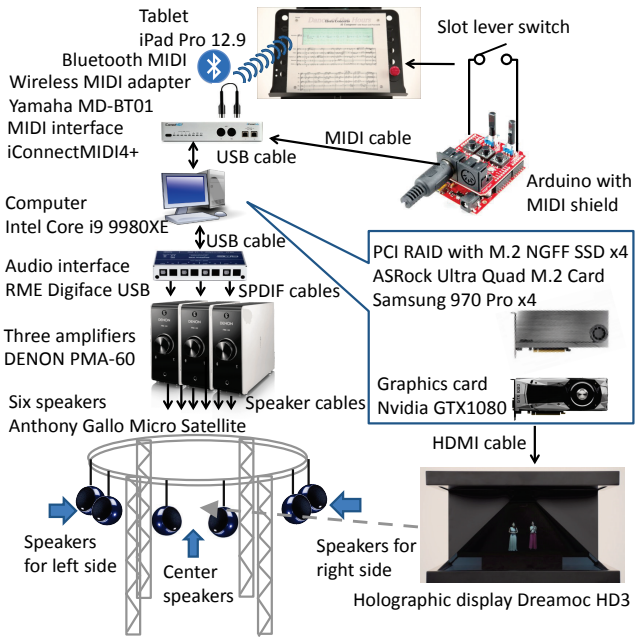


Fig. 11. Hardware implementation

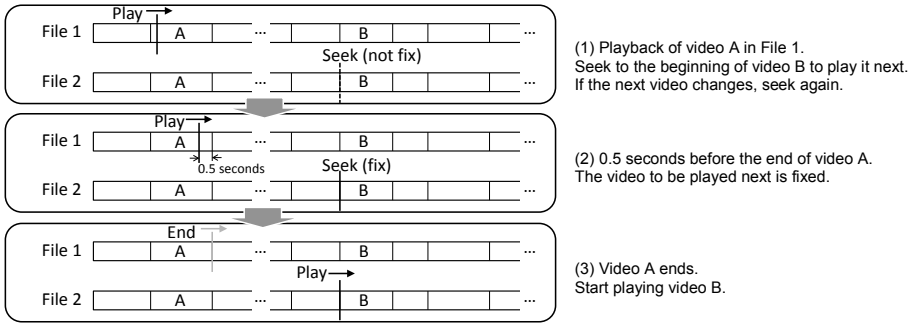


Fig. 12. Video processing

7 Experimental Results

With VIDDULL, we can set the cache size of the buffer to be read ahead when seeking the play back position. Table 1 shows the frame rate of the video when cache size changes. We used a cache size of 0.5 gigabyte, where the frame rate did not decrease much. The seeking of the playback position starts 0.5 seconds before playback. This is because with a cache size of 0.5 gigabytes, the time to seek is within 0.5 seconds. If the seeking time is less than 0.5 seconds, frame dropping occurs.

Table 1. Maximum and minimum frame rates when cache size changes

Cache size	Maximum frame rate	Minimum frame rate
0.01 gigabyte	27 fps	24 fps
0.05 gigabyte	27 fps	24 fps
0.08 gigabyte	28 fps	24 fps
0.15 gigabyte	28 fps	25 fps
0.50 gigabyte	30 fps	28 fps

8 Conclusion

In this paper, we described the Melody Slot Machine, which enables control of virtual performers on a holographic display. We plan to create various contents for the Melody Slot Machine in future works. We also plan to achieve various melody manipulations using time-span trees.

We plan to demonstrate the Melody Slot Machine and conduct a subject experiment.

Acknowledgments. This work was supported by JSPS KAKENHI, Grant Numbers 17H01847 and 16H01744.

References

1. Lerdahl, F. and Jackendoff, R.: A Generative Theory of Tonal Music. MIT Press, Cambridge, 1985.
2. Hamanaka, M., Hirata, K., and Tojo, S.: Implementing "A generative theory of tonal music." Journal of New Music Research, Vol. 35, No. 4, pp. 249–277, 2007.
3. Hamanaka, M., Hirata, K., and Tojo, S.: ATTA: Automatic Time-Span Tree Analyzer Based on Extended GTTM. Proceedings of the 6th International Conference on Music Information Retrieval Conference (ISMIR 2005), pp. 358–365, 2005.
4. Hamanaka, M., Hirata, K., and Tojo, S.: σGTTM III: Learning-Based Time-Span Tree Generator Based on PCFG. Proceedings of International Symposium on Computer Music Multidisciplinary Research (CMMR2015), pp. 387–404, 2015.
5. Hamanaka, M., Hirata, K., and Tojo, S.: deepGTTM-I: Local Boundaries Analyzer Based on Deep Learning Technique. Proceedings of International Symposium on Computer Music Multidisciplinary Research (CMMR2016), pp. 8–20, 2016.
6. Hamanaka, M., Hirata, K., and Tojo, S.: Melody Morphing Method Based on GTTM. Proceedings of International Computer Music Conference (ICMC2008), pp. 155–158, 2008.
7. Hamanaka, M., Hirata, K., and Tojo, S.: Melody Extrapolation in GTTM Approach. Proceedings of International Computer Music Conference (ICMC2009), pp. 89–92, 2009.
8. Hirata, K., and Tojo, S.: Retrograde of Melody and Flip Operation for Time-Span Tree. Proceedings of International Symposium on Computer Music Multidisciplinary Research (CMMR2016), pp. 298–305, 2016.
9. Thom, B.: Interactive Improvisational Music Companionship: A User-Modeling Approach. The User Modeling and User-Adapted Interaction Journal, Special Issue on User Modeling and Intelligent Agents, Spring 2003.
10. Wulfhorst, D. R., Nakayama, L., and Vicari, M. R.,: A multiagent approach for musical interactive systems. In AAMAS '03: Proceedings of the second international joint conference on autonomous agents and multiagent systems, pp. 584–591, 2003.
11. Assayag, G., Bloch, G., Chemillier, M., Cont, A., and Dubnov, S.: OMax brothers: a dynamic topology of agents for improvization learning. In AMCMM '06: Proceedings of the 1st ACM workshop on audio and music computing multimedia, pp. 125–132, 2006.
12. Pachet, F. The Continuator: Musical interaction with style. Journal of New Music Research, Vol. 32, No. 3, pp. 333–341, 2003.
13. Hamanaka, M., Goto, M., Asoh, H., and Otsu, N.: A learning-based jam session system that imitates a player's personality model. International Joint Conference on Artificial Intelligence (IJCAI2003), vol. 18, pp. 51–58, 2003.
14. Goto, M., Hidaka, I., Matsumoto, H., Kuroda, Y., and Muraoka, Y.: A Jazz Session System for Interplay among All Players - VirJa Session (Virtual Jazz Session System) -. Proceedings of the 1996 International Computer Music Conference (ICMC1996), pp. 346–349, 1996.
15. Hirata, K., and Matsuda, S.: Interactive Music Summarization based on Generative Theory of Tonal Music, Journal of New Music Research, Vol. 32, No. 2, pp. 165–177, 2003.
16. Hirata, K., and Hiraga, R., Ha-Hi-Hun plays Chopin's Etude. In Working Notes of IJCAI-03 Workshop on methods for automatic music performance and their applications in a public rendering contest, 2 pages, 2003.
17. Hamanaka, M., Hirata, K., and Tojo, S.: Melody Expectation Method Based on GTTM and TPS. Proceedings of the 9th International Conference on Music Information Retrieval Conference (ISMIR 2008), pp. 107–112, 2008.
18. Hedges, A. S: Dice Music in the Eighteenth Century. Music & Letters, Vol. 59, No. 2, pp. 180–187, 1978.
19. Hirata, K., and Aoyagi, T.: Computational Music Representation Based on the Generative Theory of Tonal Music and the Deductive Object-Oriented Database. Computer Music Journal, Vol. 27, No. 3, pp. 73–89, 2003.

MUSICNTWRK: data tools for music theory, analysis and composition

Marco Buongiorno Nardelli[1,2,3,4,5][0000−0003−0793−5055]

[1] CEMI, Center for Experimental Music and Intermedia, University of North Texas, Denton, TX 76203, USA
[2] iARTA, Initiative for Advanced Research in Technology and the Arts, University of North Texas, Denton, TX 76203, USA
[3] Department of Physics, University of North Texas, Denton, TX 76203, USA
[4] IMéRA - Institut Méditerranen de Recherche Avancée of Aix-Marseille Université, Marseille 13004, France
[5] Laboratoire CNRS-PRISM, 13402 Marseille, France
mbn@unt.edu
http://www.musicntwrk.com

Abstract. We present the API for MUSICNTWRK, a python library for pitch class set and rhythmic sequences classification and manipulation, the generation of networks in generalized music and sound spaces, deep learning algorithms for timbre recognition, and the sonification of arbitrary data. The software is freely available under GPL 3.0 and can be downloaded at www.musicntwrk.com.

Keywords: Computational Music Theory · Computer Aided Composition · Data Tools · Machine Learning.

1 Introduction

Big data tools have become pervasive in virtually every aspects of culture and society. In music, application of such techniques in Music Information Retrieval applications are common and well documented. However, a full approach to musical analysis and composition is not yet available for the music community and there is a need for providing a more general education on the potential, and the limitations, of such approaches. From a more fundamental point of view, the abstraction of musical structures (notes, melodies, chords, harmonic or rhythmic progressions, timbre, etc.) as mathematical objects in a geometrical space is one of the great accomplishments of contemporary music theory. Building on this foundation, we have generalized the concept of musical spaces as networks and derive functional principles of compositional design by the direct analysis of the network topology. This approach provides a novel framework for the analysis and quantification of similarity of musical objects and structures, and suggests a way to relate such measures to the human perception of different musical entities. The original contribution of this work is in the introduction of the representation of musical spaces as large-scale statistical mechanics networks:

uncovering their topological structure is a fundamental step to understand their underlying organizing principles, and to unveil how classifications or rule-based frameworks (such as common-practice harmony, for instance) can be interpreted as emerging phenomena in a complex network system. Results from this research and the theoretical and technical foundation for this paper can be found in Ref. [1].

This paper is intended to introduce the community of computer music practitioners, composers and theorists to the practical use of data science tools (network theory, machine learning etc.) through the `MUSICNTWRK` package (www.musicntwrk.com), a python library comprised of four modules:

1. `pcsPy` - pitch class set classification and manipulation; construction of generalized pitch class set networks using distances between common descriptors; the analysis of scores and the generation of compositional frameworks;
2. `rhythmPy` - rhythmic sequence classification and manipulation; and construction of rhythmic sequence networks using various definitions of rhythmic distance;
3. `timbrePy` - orchestration color networks; analysis and characterization of timbre from a (psycho-)acoustical point of view; and machine learning models for timbre recognition; and
4. `sonifiPy` - a module for the sonification of arbitrary data structures, including automatic score (musicxml) and MIDI generation.

In the following we will discuss the API of each module. All theoretical and technical analysis, including the definition of all the quantities that `MUSICNTWRK` is able to calculate, are not explicitly discussed here. The interested reader can find all this information in Ref. [1].

2 MUSICNTWRK

`MUSICNTWRK` is a python library written by the author and available at https://www.musicntwrk.com or on GitHub: https://github.com/marcobn/musicntwrk. `MUSICNTWRK` is written in python 3 and requires installation of the following modules via the "pip install" command:[§]

1. System modules: `sys, re, time, os`
2. Math modules: `numpy, itertools, fractions, gcd, functools`
3. Data modules: `pandas, sklearn, networkx, community, tensorflow`
4. Music and audio modules: `music21, librosa`
5. Parallel processing: `mpi4py`
6. Visualization modules: `matplotlib, vpython` (optional)

The reader is encouraged to consult the documentation of each package to get acquainted with its purposes and use. In what follows we provide the full API of `MUSICNTWRK` only.

[§] this step is unnecessary if running on a cloud service like Google Colaboratory.

2.1 pcsPy

pcsPy is a module for pitch class set classification and manipulation in any arbitrary temperament; the construction of generalized pitch class set networks using distances between common descriptors (interval vectors, voice leadings); the analysis of scores and the generation of compositional frameworks.

The PCSet class. pcsPy is is comprised of the PCSet class and its methods (listed below) and a series of functions for pcs network manipulations. The PCSet class deals with the classification and manipulation of pitch set classes generalized to arbitrary temperament systems (arbitrary number of pitches). The following methods are available:

`def class PCSet`

- `def __init__(self,pcs,TET=12,UNI=True,ORD=True)`
  - `pcs (int)` pitch class set as list or numpy array
  - `TET (int)` number of allowed pitches in the totality of the musical space (temperament). Default = 12 tones equal temperament
  - `UNI (logical)` if True, eliminate duplicate pitches (default)
  - `ORD (logical)` if True, sorts the pcs in ascending order (default)
- `def normalOrder(self)`
  Order the pcs according to the most compact ascending scale in pitch-class space that spans less than an octave by cycling permutations.
- `def normal0Order(self)`
  As normal order, transposed so that the first pitch is 0
- `def transpose(self,t=0)`
  Transposition by t (int) units (modulo TET)
- `def zeroOrder(self)`
  transposed so that the first pitch is 0
- `def inverse(self)`
  inverse operation: (-pcs modulo TET)
- `def primeForm(self)`
  most compact normal 0 order between pcs and its inverse
- `def intervalVector(self)`
  total interval content of the pcs
- `def LISVector(self)`
  Linear Interval Sequence Vector: sequence of intervals in an ordered pcs
- `def operator(self,name)`
  operate on the pcs with a distance operator
  - `name (str)` name of the operator O(ni)
- `def forteClass(self)`
  Name of pcs according to the Forte classification scheme (only for TET=12)
- `def jazzChord(self)`
  Name of pcs as chord in a jazz chart (only for TET=12 and cardinalities 7)
- `def commonName(self)`
  Display common name of pcs (music21 function - only for TET=12)

- `def commonNamePrime(self)`
  As above, for prime forms
- `def nameWithPitchOrd(self)`
  Name of chord with first pitch of pcs in normal order
- `def nameWithPitch(self)`
  Name of chord with first pitch of pcs
- `def displayNotes(self,xml=False,prime=False)`
  Display pcs in score in musicxml format. If prime is True, display the prime form.
  - `xml (logical)` write notes on file in musicxml format
  - `prime (logical)` write pcs in prime form

Network functions. `pcsPy` contains specific functions for network generation and analysis. Network functions include:

- `def pcsDictionary(Nc,order=0,TET=12,row=False,a=np.array(None))`
  Generate the dictionary of all possible pcs of a given cardinality in a generalized musical space of TET pitches. Returns the dictionary as pandas DataFrame and the list of all Z-related pcs
  - `Nc (int)` cardinality
  - `order (logical)` if 0 returns pcs in prime form, if 1 retrns pcs in normal order, if 2, returns pcs in normal 0 order
  - `row (logical)` if True build dictionary from tone row, if False, build dictionary from all combinatorial pcs of Nc cardinality given the totality of TET.
  - `a (int)` if row = True, a is the list of pitches in the tone row
- `def pcsNetwork(input_csv, thup=1.5, thdw=0.0,TET=12, distance= 'euclidean', col=2,prob=1)`
  generate the network of pcs based on distances between interval vectors
  In output it writes the nodes.csv and edges.csv as separate files in csv format
  - `input_csv (str)` file containing the dictionary generated by pcsNetwork
  - `thup, thdw (float)` upper and lower thresholds for edge creation
  - `distance (str)` choice of norm in the musical space, default is 'euclidean'
  - `col (int)` metric based on interval vector, col = 1 can be used for voice leading networks in spaces of fixed cardinality NOT RECOMMENDED
  - `prob (float)` if not 1, defines the probability of acceptance of any given edge
- `def pcsEgoNetwork(label,input_csv,thup_e=5.0,thdw_e=0.1,thup=1.5,thdw=0.1, TET=12,distance='euclidean')`
  Generates the network for a focal node (ego) and the nodes to whom ego is directly connected to (alters). In output it writes the nodes_ego.csv, edges_ego.csv and edges_alters.csv as separate files in csv format
  - `label (str)` label of the ego node
  - `thup_e, thdw_e (float)` upper and lower thresholds for edge creation from ego node

  - `thup, thdw (float)` upper and lower thresholds for edge creation among alters
  - `distance (str)` choice of norm in the musical space, default is 'euclidean'
- `def vLeadNetwork(input_csv,thup=1.5,thdw=0.1,TET=12,w=True, distance='euclidean',prob=1)`
  Generation of the network of all minimal voice leadings in a generalized musical space of TET pitches based on the minimal distance operators select by distance. In output returns nodes and edges tables as pandas DataFrames.
  - `input_csv (str)` file containing the dictionary generated by pcsNetwork
  - `thup, thdw (float)` upper and lower thresholds for edge creation
  - `distance (str)` choice of norm in the musical space, default is 'euclidean'
  - `w (logical)` if True it writes the nodes.csv and edges.csv files in csv format
- `def vLeadNetworkByName(input_csv,thup=1.5,thdw=0.1,TET=12,w=True, distance='euclidean',prob=1)`
  Generation of the network of all minimal voice leadings in a generalized musical space of TET pitches based on the minimal distance operators select by name. In output returns nodes and edges tables as pandas DataFrames. Available also in vector form for computational efficiency as `vLeadNetworkByNameVec`
  - `input_csv (str)` file containing the dictionary generated by pcsNetwork
  - `name (str)` name of operator for edge creation
  - `distance (str)` choice of norm in the musical space, default is 'euclidean'
  - `w (logical)` if True it writes the nodes.csv and edges.csv files in csv format
- `def scoreNetwork(seq,TET=12)`
  Generates the directional network of chord progressions from any score in musicxml format
  - `seq (int)` list of pcs for each chords extracted from the score
- `def scoreDictionary(seq,TET=12)`
  Builds the dictionary of pcs in any score in musicxml format
- `def readScore(inputxml,TET=12,music21=False)`
  Reads musicxml score and returns chord sequence
  - `inputxml (str)` score file
  - `music21 (logical)` if True search the music21 corpus

2.2 `rhythmPy`.

`rhythmPy` is a module for rhythmic sequence classification and manipulation; and the construction of rhythmic sequence networks using various definitions of rhythmic distance.

The ***RHYTHMSeq*** *class* `rhythmPy` is comprised of the `RHYTHMSeq` class and its methods (listed below) and a series of functions for rhythmic network manipulations. The `RHYTHMSeq` class deals with the classification and manipulation of rhythmic sequences. The following methods are available:

`def class RHYTHMSeq`

- `__init__( self,rseq,REF='e',ORD=False)`
  - `rseq (str/fractions/floats)` rhythm sequence as list of strings/fractions/floats name
  - `REF (str)` reference duration for prime form the RHYTHMSeq class contains a dictionary of common duration notes that uses the fraction module for the definitions (implies import fraction as fr): {`'w':fr.Fraction(1,1), 'h':fr.Fraction(1,2),'q':fr.Fraction(1,4), 'e':fr.Fraction(1,8), 's':fr.Fraction(1/16),'t':fr.Fraction(1,32), 'wd':fr.Fraction(3,2), 'hd':fr.Fraction(3,4),'qd':fr.Fraction(3,8), 'ed':fr.Fraction(3,16), 'sd':fr.Fraction(3,32),'qt':fr.Fraction(1,6), 'et':fr.Fraction(1,12), 'st':fr.Fraction(1,24), 'qq':fr.Fraction(1,5), 'eq':fr.Fraction(1,10), 'sq':fr.Fraction(1,20)`}. This dictionary can be extended by the user on a case by case need.
  - `ORD (logical)` if `True` sort durations in ascending order
- `def normalOrder(self)`
  Order the rhythmic sequence according to the most compact ascending form.
- `def augment(self,t='e')`
  Augmentation by t units
  - `t (str` duration of augmentation
- `def diminish(self,t='e')`
  Diminution by t units
  - `t (str)` duration of diminution
- `def retrograde(self)`
  Retrograde operation
- `def isNonRetro(self)`
  Check if the sequence is not retrogradable
- `def primeForm(self)`
  reduce the series of fractions to prime form
- `def durationVector(self,lseq=None)`
  total relative duration ratios content of the sequence
  - `lseq (list of fractions)` reference list of duration for evaluating interval content the default list is: [fr.Fraction(1/8),fr.Fraction(2/8), fr.Fraction(3/8), fr.Fraction(4/8),fr.Fraction(5/8), fr.Fraction(6/8), fr.Fraction(7/8), fr.Fraction(8/8), fr.Fraction(9/8)]
- `def durationVector(self,lseq=None)`
  inter-onset duration interval content of the sequence
  - `lseq (list of fractions)` reference list of duration for evaluating interval content the default list is the same as above.

Network functions. `rhythmPy` contains specific functions for network generation and analysis. Network functions include:

- `def rhythmDictionary(Nc,a=None,REF='e')`
  Generates the dictionary of all possible rhythmic sequences of Nc length in a generalized meter space of N durations. Returns the dictionary as pandas DataFrame and indicates all non retrogradable and Z-related cells
  - `Nc (int)` cell length
  - `a (str)` list of durations in the rhythm sequence
- `def rhythmPDictionary(N,Nc,REF='e')`
  Generate the dictionary of all possible rhythmic sequences from all possible groupings of N REF durations. Returns the dictionary as pandas DataFrame and indicates all non retrogradable and Z-related cells
  - `Nc (int)` cell length
  - `N (int)` number of REF units
- `def rhythmNetwork(input_csv,thup=1.5,thdw=0.0,distance='euclidean',prob=1,w=False)`
  Generates the network of rhythmic cells based on distances between duration vectors. In output it writes the nodes.csv and edges.csv as separate files in csv format
  - `input_csv (str)` file containing the dictionary generated by rhythmNetwork
  - `thup, thdw (float)` upper and lower thresholds for edge creation
  - `distance (str)` choice of norm in the musical space, default is 'euclidean'
  - `prob (float)` if not 1, defines the probability of acceptance of any given edge
  - `w (logical)` if True it writes the nodes.csv and edges.csv files in csv format
- `def rLeadNetwork(input_csv,thup=1.5,thdw=0.1,w=True,distance='euclidean',prob=1)`
  Generation of the network of all minimal rhythm leadings in a generalized musical space of Nc-dim rhythmic cells based on the rhythm distance operator. Returns nodes and edges tables as pandas DataFrames
  - `input_csv (str)` file containing the dictionary generated by rhythmNetwork
  - `thup, thdw (float)` upper and lower thresholds for edge creation
  - `distance (str)` choice of norm in the musical space, default is 'euclidean'
  - `prob (float)` if not 1, defines the probability of acceptance of any given edge
  - `w (logical)` if True it writes the nodes.csv and edges.csv files in csv format

2.3 `timbrePy`

`timbrePy` comprises of two sections: the first deals with orchestration color and it is the natural extension of the score analyzer in `pscPy`; the second deals with analysis and characterization of timbre from a (psycho-)acoustical point of view. In particular, it provides: the characterization of sound using, among others, Mel Frequency or Power Spectrum Cepstrum Coefficients (MFCC or PSCC); the construction of timbral networks using descriptors based on MF- or PS-CCs; and machine learning models for timbre recognition through the TensorFlow Keras framework.

Orchestration analysis. The orchestration analysis section of `timbrePy` comprises of the following modules:

- `def orchestralVector(inputfile,barplot=True)`
  Builds the orchestral vector sequence from score in `musicxml` format. Returns the score sliced by beat; orchestration vector.
  - `barplot=True` plot the orchestral vector sequence as a matrix
- `def orchestralNetwork(seq)`
  Generates the directional network of orchestration vectors from any score in musicxml format. Use orchestralScore() to import the score data as sequence. Returns nodes and edges as Pandas DataFrames; average degree, modularity and partitioning of the network.
  - `seq (int)` list of orchestration vectors extracted from the score
- `def orchestralVectorColor(orch,dnodes,part,color=plt.cm.binary)`
  Plots the sequence of the orchestration vectors color-coded according to the modularity class they belong. Requires the output of orchestralNetwork()
  - `seq (int)` list of orchestration vectors extracted from the score

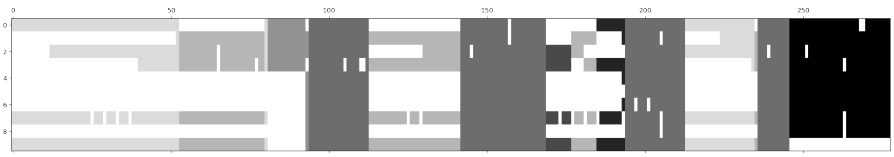


Fig. 1. Orchestration map of the first movement (Allegro) of J.S. Bach's Brandenburg Concerto n. 2, BWV 1047, as produced by the `orchestralVectorColor` function. Different shades of gray represent the sections of similar orchestral color as measured by their modularity class in the network.

Sound classification. The sound classification section of `timbrePy` comprises of modules for specific sound analysis that are based on the `librosa` python library for audio signal processing. We refer the interested reader to the `librosa` documentation at `https://librosa.github.io/librosa/index.html`. Fore a more complete discussion on the descriptors defined in `MUSICNTWRK` please refer to the work in Ref. [2]. Specific audio signal processing functions are:

- `def computeMFCC(input_path,input_file,barplot=True,zero=True)`
  read audio files in repository and compute a normalized MEL Frequency Cepstrum Coefficients and single vector map of the full temporal evolution of the sound as the convolution of the timeresolved MFCCs convoluted with the normalized first MFCC component (power distribution). Returns the list of files in repository, MFCC0, MFCC coefficients.
  - `input_path (str)` path to repository
  - `input_file (str)` filenames (accepts "*")
  - `barplot (logical)` plot the MFCC0 vectors for every sound in the repository
  - `zero (logical)` If False, disregard the power distribution component.
- `def computePSCC(input_path,input_file,barplot=True,zero=True)`
  Reads audio files in repository and compute a normalized Power Spectrum Frequency Cepstrum Coefficients and single vector map of the full temporal evolution of the sound as the convolution of the time-resolved PSCCs convoluted with the normalized first PSCC component (power distribution). Returns the list of files in repository, PSCC0, PSCC coefficients. Other variables as above.
- `def computeStandardizedMFCC(input_path,input_file,nmel=16, nmfcc=13,lmax=None,nbins=None)`
  read audio files in repository and compute the standardized (equal number of samples per file) and normalized MEL Frequency Cepstrum Coefficient. Returns the list of files in repository, MFCC coefficients, standardized sample length.
  - `nmel (int)` number of Mel bands to use in filtering
  - `nmfcc (int)` number of MFCCs to return
  - `lmax (int)` max number of samples per file
  - `nbins (int)` number of FFT bins
- `def computeStandardizedPSCC(input_path,input_file,nmel=16, psfcc=13,lmax=None,nbins=None)`
  read audio files in repository and compute the standardized (equal number of samples per file) and normalized Power Spectrum Frequency Cepstrum Coefficients. Returns the list of files in repository, PSCC coefficients, standardized sample length.
  Variables defined as for MFCCs.
- `def timbralNetwork(waves,vector,thup=10,thdw=0.1)`
  generates the network of MFCC vectors from sound recordings. Returns the nodes and edges tables as pandas DataFrames
  - `seq (float)` list of MFCC0 vectors
  - `waves (str)` names of sound files

Machine Learning Models. The definition of machine learning models for sound recognition requires standard techniques of data science (like the separation of data entries in training and testing sets, definition of neural network architectures, etc.) that will not be discussed here. Basic knowledge of Keras is also

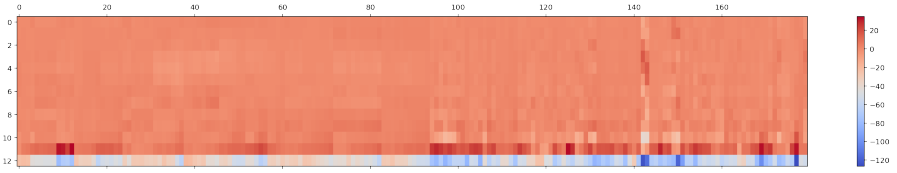


Fig. 2. Map of the MFCC0 for a repository of 180 impact sounds.

assumed. `MUSICNTWRK` module `timbrePy` contains many auxiliary functions to deal with such tasks. Here we limit to report the API for the main machine learning functions:

- `def trainNNmodel(mfcc,label,gpu=0,cpu=4,niter=100,nstep=10, neur=16,test=0.08,num_classes=2,epoch=30,verb=0,thr=0.85,w=False)` train a 2 layer neural network model on the ful MFCC spectrum of sounds. Returns: model,training and testing sets,data for re-scaling and normalization,data to asses the accuracy of the training session.
  - `mfcc (float)` list of all the MFCCs (or PSCCs) in the repository
  - `gpu, cpu (int)` number of GPUs or CPSs used for the run
  - `niter (int)` max number of model fit sessions
  - `nstep (int)` how often the training and testing sets are redefined
  - `neur (int)` number of neurons in first layer (it is doubled on the second layer
  - `test (float)` defines the relative size of training and testing sets
  - `num_classes=2 (int)` dimension of the last layer
  - `epoch (int)` number of epochs in the training of the neural network
  - `verb (int)` verbose - print information during the training run
  - `thr (float)` keep the model if accuracy is ¿ test
  - `w (logical)` write model on file if accuracy is above `thr`
- `def trainCNNmodel(mfcc,label,gpu=0,cpu=4,niter=100,nstep=10, neur=16,test=0.08,num_classes=2,epoch=30,verb=0,thr=0.85,w=False)` train a convolutional neural network (CNN) model on the full MFCC/PSCC spectrum of sounds. Returns: model,training and testing sets,data for re-scaling and normalization,data to asses the accuracy of the training session. Parameters are defined as above.

2.4 `sonifiPy`

`sonifiPy` contains functions for the sonification of data in multi-column or csv format and produces output as WAV (it requires an installation of csound and direct reference to the `ctcsound` module -), or musicxml or MIDI. Two sonification protocols are available: spectral - data are mapped to a single sound using subtractive synthesis (FIR filter); and linear - individual data points are mapped to pitches in a time-series structure. See Ref. [3, 4] for a complete description of this protocol. `sonifiPy` contains:

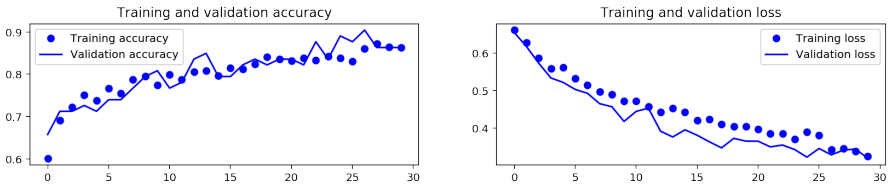


Fig. 3. Training and validation accuracy and loss in a typical Neural Network model learning run

- `def r_1Ddata(path,fileread)`
  Read data file in a multicolumn format (csv files can be easily put in this format using Pandas). Returns the data values as (x,y).
  - `path (str)` path to data file
  - `fileread (str)` data file
- `def i_spectral2(xv,yv,itime,path='./',instr='noise')`
  Use subtractive synthesis to sonify data structure. Returns the sound file.
  - `xv,yv (float)` data structure to sonify
  - `path (str)` path to data file
  - `fileread (str)` data file
- `def i_time_series(xv,yv,path='./',instr='csb701')`
  Use csound instruments to sonify data structures as time-series. Returns the sound file.
  - `xv,yv (float)` data structure to sonify
  - `path (str)` path to data file
  - `fileread (str)` data file
  - `instr (str)` csound instrument (it can be modified by user)
- `def MIDImap(pdt,scale,nnote)`
  Data to MIDI conversion on a given scale defined in scaleMapping (see below). Returns the MIDI data structure.
  - `pdt (float)` data structure mapped to MIDI numbers
  - `scale (float)` scale mapping (from scaleMapping)
  - `nnote (int)` number of notes in the scale (from scaleMapping)
- `def scaleMapping(scale)`
  Scale definitions for MIDI mapping. Returns: scale, nnote (see above).
- `def MIDIscore(yvf,dur=2,w=None,outxml='./music',outmidi='./music')`
  Display score or writes to file
  - `yvf (float)` data structure mapped to MIDI numbers (from MIDImap)
  - `dur (int)` reference duration
  - `w (logical)` if True writes either musicxml or MIDI file)
- `def MIDImidi(yvf,vnorm=80,dur=4,outmidi='./music')`
  Display score or writes to file
  - `yvf (float)` data structure mapped to MIDI numbers (from MIDImap)
  - `vnorm (int)` reference velocity
  - `outmidi (str)` MIDI file

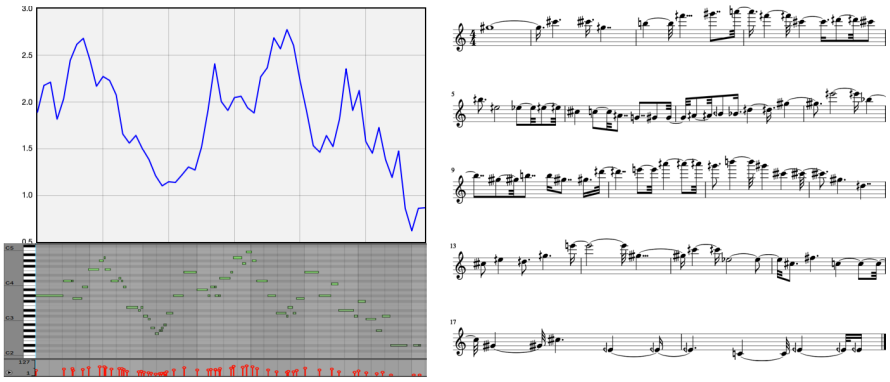


Fig. 4. Data, MIDI map and score from the sonification protocol in `MIDIscore`

The most computationally intensive parts of the modules can be run on parallel processors using the MPI (Message Passing Interface) protocol. Communications are handled by two additional modules: `communications` and `load_balancing`. Since the user will never have to interact with these modules, we omit here a detailed description of their functions.

Finally, a full set of examples is provided as Jupyter notebooks with the distribution.

3 Conclusions and acknowledgments

We have presented the API for the `MUSICNTWRK` software package. The software is freely available under GPL 3.0 and can be downloaded at www.musicntwrk.com. We acknowledge the support of Aix-Marseille University, IMéRA, and of Labex RFIEA+. Finally, we thank Richard Kronland-Martinet, Sølvi Ystad, Mitsuko Aramaki, Jon Nelson, Joseph Klein, Scot Gresham-Lancaster, David Bard-Schwarz, Roger Malina and Alexander Veremyer for useful discussions.

References

1. Buongiorno Nardelli, M.: Topology of Networks in Generalized Musical Spaces. https://arxiv.org/abs/1905.01842, 2019.
2. Buongiorno Nardelli, M., Aramaki, M., Ystad, S., Kronland-Martinet, R.: *in preparation*, 2019
3. Buongiorno Nardelli, M.: materialssoundmusic: a computer-aided data-driven composition environment for the sonification and dramatization of scientific data streams. International Computer Music Conference Proceedings, **356** (2015).
4. Buongiorno Nardelli, M.: Beautiful Data: Reflections for a Sonification and Post-Sonification Aesthetics, in Leonardo Gallery: Scientific Delirium Madness 4.0, Leonardo **51**(3), 227–238 (2018).

Feasibility Study of Deep Frequency Modulation Synthesis

Keiji Hirata[1⋆], Masatoshi Hamanaka[2], and Satoshi Tojo[3]

[1] Future University Hakodate hirata@fun.ac.jp
[2] Riken AIP masatoshi.hamanaka@riken.jp
[3] JAIST tojo@jaist.ac.jp

Abstract. Deep Frequency Modulation (FM) synthesis is the method of generating approximate or new waveforms by the network inspired by the conventional FM synthesis. The features of the method include that the activation functions of the network are all vibrating ones with distinct parameters and every activation function (oscillator unit) shares an identical time t. The network learns a training waveform given in the temporal interval designated by time t and generates an approximating waveform in the interval. As the first step of the feasibility study, we examine the basic performances and potential of the deep FM synthesis in small-sized experiments. We have confirmed that the optimization techniques developed for the conventional neural networks is applicable to the deep FM synthesis in small-sized experiments.

Keywords: Frequency modulation synthesis · neural networks · activation function · backpropagation.

1 Introduction

Frequency Modulation (FM) synthesis is a well-known technique for generating musical sound [1] and has been employed for many commercial products of digital synthesizers such as DX7 of Yamaha, which is one of the bestselling synthesizers [7]. Also many variations of the FM synthesis have been developed [6, pp.224-250]). FM synthesis can generate rich sounds despite a simple configuration, i.e., the small number of parameters; on the other hand, it is known that FM synthesis requires some skills for manipulating parameters when generating new desired sounds. It is mainly because the relationships among output sounds, parameter values, and configurations of connecting oscillators are not sufficiently intuitive. Hence, to create new sounds easily, many of digital synthesizers employing the FM synthesis provides the presets which are built-in connection patterns of oscillators with predefined parameters of amplitudes and carrier and modulating frequencies.

After considering these presets provided, we would come up with an idea of a general form of the FM synthesis which looks like a neural network, the activation functions of which are oscillators. Suppose, in the network, an oscillator

⋆ This work has been supported by JSPS Kakenhi 16H01744.

X at a layer can receive modulating waveforms from the ones at one-level lower layer $Y_1, Y_2, \cdots$. The weight of the connection between X with Y_k corresponds to amplitude parameters. The carrier and modulating frequencies are defined as parameters within each oscillator. Then, all the oscillators should refer to an identical current time and simultaneously generate waveforms along with the time as in the conventional FM synthesis. If we would apply the learning techniques developed for conventional neural networks to the network inspired by the conventional FM synthesis, we might generate target sounds without taking care of the relationships among output sounds, parameter values, and configurations of connecting oscillators.

Gashler and Ashmore [2] have surveyed various networks to model and predict time-series data and offered a useful idea for categorization of approaches. Gashler and Ashmore claim that at a high level, the neural networks to predict time-series data are broadly categorized into three major approaches: here we refer to them as WaveNet approach, RNN approach, and extrapolation approach. Among them, there have been proposed several neural networks that belong to the extrapolation approach and employ vibrating activation functions such as a sinusoidal function and wavelet [2, 5, 3]. In these architectures, the current time for generating a waveform is treated as the explicit input given at an input layer. That is, neither all the activation functions (i.e., oscillators) share the current time, nor the activation functions at the hidden layers refer to it.

In the paper, we propose deep Frequency Modulation (FM) synthesis, which is the method of generating an approximating waveform based on the network inspired by the conventional FM synthesis. According to the Gashler and Ashmore's categorization, the deep FM synthesis basically belongs to the extrapolation approach. Thus, we hope the deep FM sysntesis could generate unknown yet good sounds by extrapolating already existing sounds in a different way from prevalent sound generation methods such as sampling and WaveNet. To study the feasibility and utility of the deep FM synthesis, we investigate the basic characteristics and performances of it; for instance, how accurate the deep FM synthesis can approximate a target waveform, what size of the network we need for reconstructing a target waveform, how and what conventional techniques for optimizing networks can be applied to the deep FM synthesis, and so on.

2 Deep Frequency Modulation Synthesis

2.1 Architecture

For theoretical consideration, we think of a simple, typical architecture shown in Figure 1, which presents how oscillator units are interconnected with weights w; the depth is $M+1$, the width is $N+1$, and Σ^w stands for weighted sum. The input to the network is vector $\{A_0, A_1, \cdots A_N\}$, and the output is waveform S. For an oscillator unit by a typical vibrating function, we here adopt a sinusoidal function $y = \sin\ 2\pi((x+c)\ t+p)$ with input x (the bottom of the oscillator unit in the figure) and output y (the top). Each oscillator unit has two parameters c and p to be tuned corresponding to frequency and phase, respectively. All

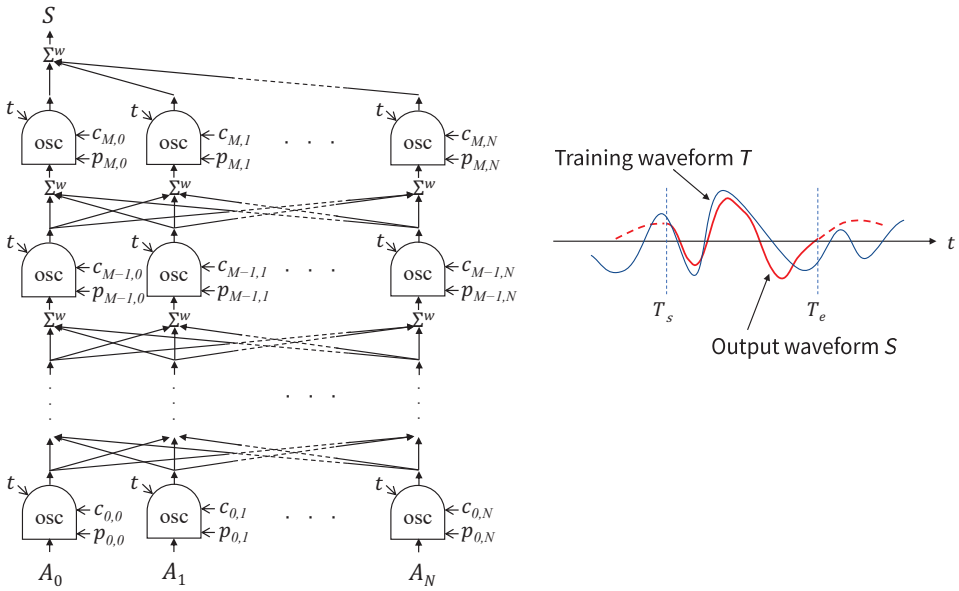


Fig. 1. Network Configuration of Deep FM Synthesis

oscillator units share the identical timing signal t, which moves between starting time T_s and ending time T_e, to compute the output waveform (the red solid curve in Figure 1). The network attempts to fit the output waveform to the training waveform only between T_s and T_e (dark blue). Thus, in the ranges out of the interval between T_s and T_e, the network does not take care of the output waveform (red dashed curves).

The forward propagation in the deep FM synthesis works as conventional neural networks; at layer n, input waveform x is given to each oscillator unit to compute output y with parameters $c_{m,n}$ and $p_{m,n}$ at time t. Then, the output waveforms calculated at layer n, that is y, are summed up with weights, and the sum is provided to the input waveform at layer $n+1$ as x. An output waveform is made of the series of values S_t, which represent the samples at time instant t between T_s and T_e. In other words, given time t at which we want to obtain value S_t, the network calculates S_t in a bottom-up manner.

We may assume that all the elements of the input vector $\{A_0, A_1, \cdots A_N\}$ are constant values. Theoretically, the input vector can be either constants or any waveforms synchronized by the timing signal t, since giving waveforms to the input is equivalent to the extension of the network in the direction of depth with constants given to the input. Therefore, the multi-layered architecture absorbed such subtle differences, and we can put the assumption without loss of generality.

2.2 Backpropagation

We would apply the standard backpropagation technique to optimize the deep FM synthesis as follows [4]. For notational simplification, we assume the network size is depth $M+1$ by width $N+1$, and the width is unchanged from the top layer to the bottom. The final output of the deep FM synthesis at time t is denoted as S_t. The input and output of the n-th oscillator unit at layer m are denoted as $x_{m,n}$ and $y_{m,n}$, respectively. The weight between adjacent layers is denoted as $w_{m,n',n}$, which stands for the weight from n'-th unit at layer m to n-th unit at layer $m+1$. Only at the topmost layer, we write $w_{M,n}$, omitting the second n. Then, the final output is straightforwardly defined in a topdown manner:

$$S_t = \sum_{n=0}^{N} w_{M,n} \cdot y_{M,n} \tag{1}$$

For m = M .. 1 and n = 0 .. N, we define $y_{m,n} = \sin 2\pi((x_{m,n} + c_{m,n})t + p_{m,n})$ and $x_{m,n} = \sum_{n'=0}^{N} w_{m-1,n',n} \cdot y_{m-1,n'}$. For m = 0 and n = 0 .. N (the bottom layer), we define $y_{0,n} = \sin(x_{0,n}t + p_{0,n})$, and $x_{0,n} = A_n$. We always put $c_{0,n} = 0$ because the elements of the input $\{A_0, A_1, \cdots A_N\}$ are all assigned to constant values.

A single network of deep FM synthesis is trained, considering the set of time instants within the designated period between T_s and T_e. Let us denote the training waveform (target waveform) at time t as T_t. Then, the loss function is defined as follows:

$$E = \sum_t E_t = \sum_t \frac{1}{2}(S_t - T_t)^2 \,, \tag{2}$$

where $\sum_t$ means the summation over the set of the time instants ($T_s \le t \le T_e$).

We present the gradient descent method for optimizing the network [4]; let us compute the partial differential of the loss in Equation (2) with respect to each parameter contained in the network in the standard manner. First of all, for the topmost weights $w_{M,n}$, from Equation (1) we have

$$\frac{\partial E_t}{\partial w_{M,n}} = \frac{\partial E_t}{\partial S_t} \cdot \frac{\partial S_t}{\partial w_{M,n}} = (S_t - T_t) \cdot y_{M,n}$$

Note that although the same timing signal t is provided to all oscillator units, the gradient of the loss can be derived as in conventional neural networks. Also for the parameters within each oscillator, $\frac{\partial E_t}{\partial c_{M,n}}$ and $\frac{\partial E_t}{\partial p_{M,n}}$ can be derived similarly. Then, to simply express the derivatives of $y_{m,n}$, we introduce term $\cos 2\pi((x_{m,n} + c_{m,n})t + p_{m,n})$ and denote it as $z_{m,n}$. Due to space limitation, we omit the technical details and show only the result of parameter optimization. We derive the following entire inductive definition of the gradient chain Γ and the feedback values for gradient descent:
Base step: $\Gamma^A_{M,n} = (S_t - T_t) \cdot w_{M,n}$

Induction step:

$$\begin{aligned}
\Gamma^A_{m,n} &= \sum_{i=0}^{N} \Gamma^B_{m+1,i} \cdot \ w_{m,n,i} \quad \text{(m = 0 .. M-1)} \\
\Gamma^B_{m,n} &= \Gamma^A_{m,n} \cdot 2\pi t \ z_{m,n} \quad \text{(m = 0 .. M)}
\end{aligned} \tag{3}$$

Using the series of the gradient chain above, we obtain the partial differentials of the parameters as follows:

$$\begin{aligned}
\frac{\partial E_t}{\partial w_{M,n}} &= (S_t - T_t) \cdot y_{M,n} \\
\frac{\partial E_t}{\partial w_{m,n',n}} &= \Gamma^B_{m+1,n} \cdot y_{m,n'} \quad \text{(m = 0 .. M-1)} \\
\\
\frac{\partial E_t}{\partial c_{m,n}} &= \Gamma^A_{m,n} \cdot 2\pi t \ z_{m,n} \quad \text{(m = 0 .. M)} \\
\frac{\partial E_t}{\partial p_{m,n}} &= \Gamma^A_{m,n} \cdot 2\pi \ z_{m,n} \quad \text{(m = 0 .. M)}
\end{aligned}$$

3 Experiments and Results

3.1 Implementation

Following the standard backpropagation techniques [4], for optimization, we employ stochastic gradient descent, Adam, and L^2 regularization with soft threshold. Within an epoch, as many time instants at which the loss is calculated as the size of the mini-batch are generated by the uniform random number generator over the designated temporal interval. In the following experiments, the mini-batch size is always set to 5. For simplicity, all parameters c's, p's, and w's are initialized by the normal distribution with the average being 0.0 and the variance being 0.1. Each layer, consisting of the oscillator units, is fully-connected to adjacent layers.

Here, we give a notice in setting the period of time t to preserve stability for the deep FM synthesis. As some may already noticed in Equation (3), the gradient chain inevitably includes term t^n, where n is the depth from the top layer. It follows that the amount of the loss feedback is proportional to t^n. Thus, if $0.0 < t < 1.0$, t^n may always become almost 0.0; on the other hand, if $t > 1.0$, t^n may become a larger value. Therefore, in the following experiments, the period of time t is put from 1.0 to 2.0. These values have been determined through several trials we made.

3.2 Loss Convergence

At the very first step, we would check if the backpropagation technique introduced in the previous section can work for the deep FM synthesis. Figure 2 shows

the loss convergences calculated by mean squared error as the epoch increases (Equation (2)), when the network size is depth 5 by width 5 and the training waveforms are rectangle and sawtooth with two cycles in the period of 1.0 to 2.0.

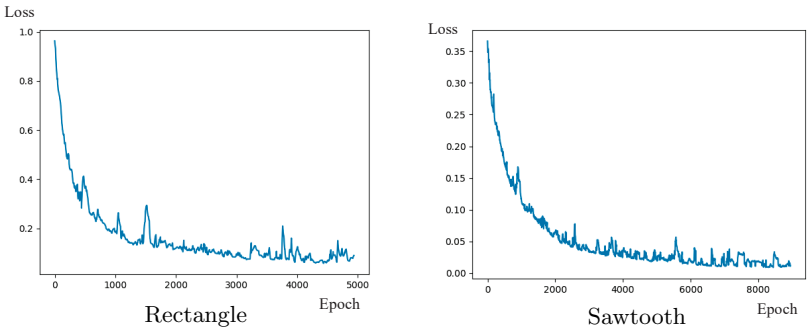


Fig. 2. Loss Convergences along with Epoch

Figure 3 shows the intermediate waveforms generated by the network at epochs 0, 500, and 4400 for rectangle and at epochs 0, 1000, and 8400 for sawtooth, respectively. Note that the training and output waveforms are shown only between T_s and T_e (i.e., 1.0 and 2.0). To obtain the above results, it took several minutes or less for each training phase, using Surface Pro 3 featuring Intel Core i7 CPU @ 1.70 GHz.

3.3 Network Size

We examine the relationship between the network size (depth × width) and the loss. The left-hand graph in Figure 4 shows the results for rectangle as the training waveform, and the right-hand graph sawtooth. In the both graphs, the depth is changed from 2 to 5 (colored broken lines), and the width from 1 to 8 (horizontal axes).

The loss in sawtooth converges faster along with the width increased than rectangle. At present, we presume the result could be understood by the complexity of a waveform as a figure. For example, while a cycle of rectangle contains two steep changes (-1.0 to 1.0 and 1.0 to -1.0), that of sawtooth contains one (1.0 to -1.0).

4 Concluding Remarks

We propose the deep FM synthesis which is inspired by the conventional FM synthesis; it has the network architecture like neural networks and can be optimized by the backpropagation technique as neural networks. We have demonstrated that the deep FM synthesis works well to some extent for small-sized

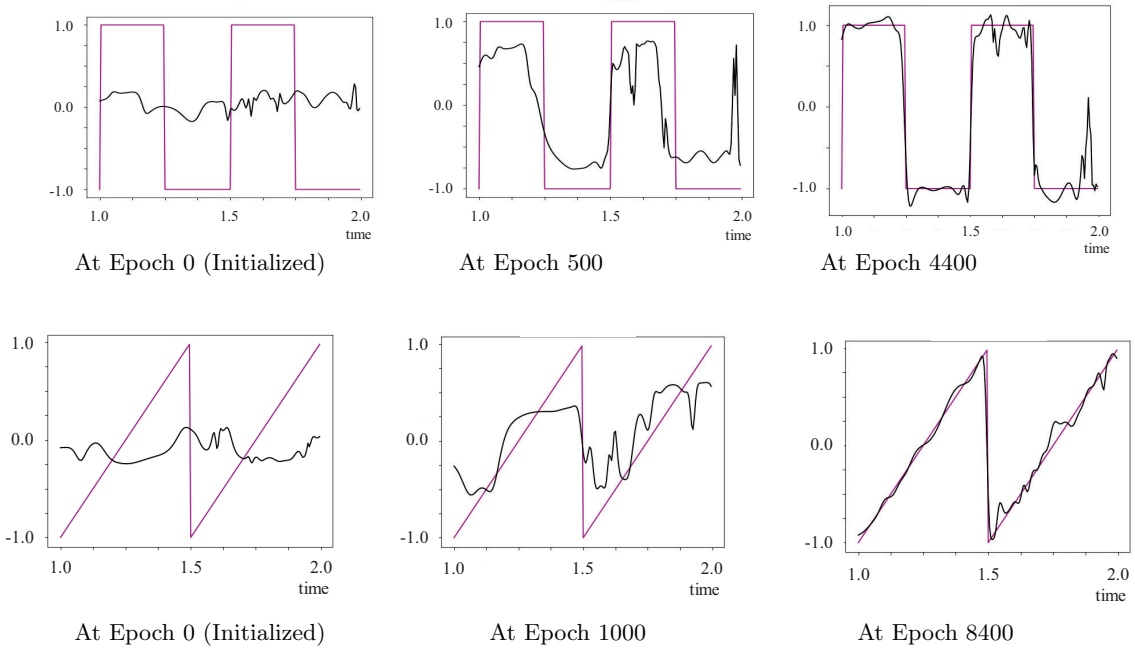


Fig. 3. Waves Generated During Training for Rectangle and Sawtooth

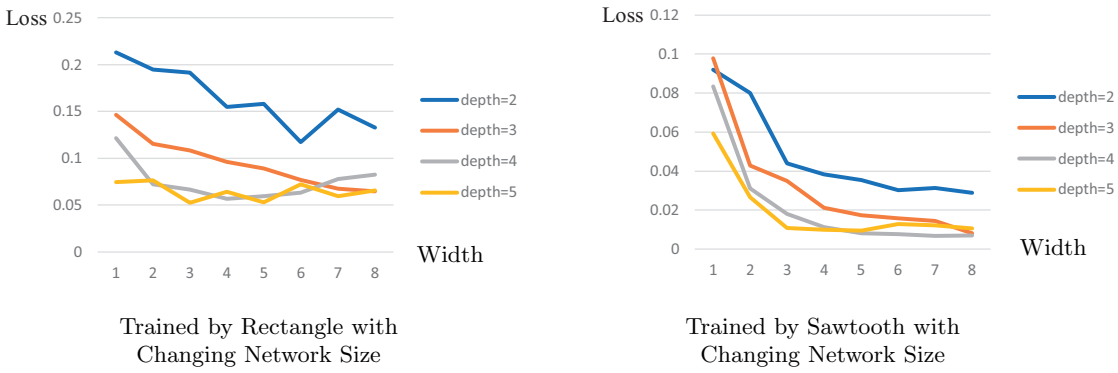


Fig. 4. Network Size and Loss

artificial training waveforms. Figure 5 shows an example when an unsteady, a little complicated training waveform (black curve in the figure) is given; the wave

length and amplitude of it is varied along time. The horizontal axis in the figure stands for time in the unit of time instant.

The network attempts to approximate the output waveform (blue) to the training waveform; the network size is here depth 4 by width 20, and the network achieves the loss of 0.0583 at epoch 7570. However, the realistic waveforms of natural tones of acoustic instruments such as pianos or flutes are far longer and contain many cycles. At present, the network of the deep FM synthesis cannot properly learn and reconstruct such real waveforms, unfortunately. The output waveform of the network does not sufficiently converge on a given training waveform under the current optimization method.

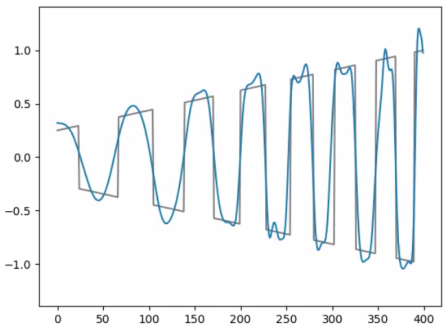


Fig. 5. Approximating Unsteady Waveform

Future work will include improving the approximation to the simple waveforms as given in Sections 3.2 and 3.3, and developing a tractable optimization method that can work effectivtly when learning realistic, long, complicated waveforms such as natural tones of acoustic instruments and voices.

References

1. Chowning, J.M.: The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. J. the Audio Engineering Society **7**(21), 526–534 (1973)
2. Gashler, M.S., Ashmore, S.C.: Training Deep Fourier Neural Networks To Fit Time-Series Data (2014), arXiv preprint arXiv:1405.2262v1 (2014)
3. Godfrey, L.B.: Parameterizing and Aggregating Activation Functions in Deep Neural Networks. Ph.D. thesis, University of Arkansas (May 2018)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press (2016)
5. Mingo, L., Aslanyan, L., Castellanos, J., Daz, M., Riazanov, V.: Fourier Neural Networks: An Approach With Sinusoidal Activation Functions. International Journal "Information Theories & Applications" **11**, 52–55 (2004)
6. Roads, C.: The Computer Music Tutorial. The MIT Press (1996)
7. Wikipedia: Yamaha DX7. https://en.wikipedia.org/wiki/Yamaha_DX7, accessed: 2019/01/24

Description of Monophonic Attacks in Reverberant Environments via Spectral Modeling

Thiago A. M. Campolina[1] and Mauricio Alves Loureiro[1] *

Center for Studies on Musical Gesture & Expression (CEGeME), School of Music,
Federal University of Minas Gerais (UFMG) - Brazil
thicampolina@gmail.com, mauricio.alves.loureiro@gmail.com

Abstract. Note attack has an important role in music performance. By manipulating note attack quality, musicians are able to control timbre, articulation, rhythm, essential parameters for conveying their expressive intentions. We explored the idea of describing content of monophonic musical note attacks via Spectral Modeling Synthesis decomposition considering, independently, the harmonic content of reverberation in the attack region of the note being played. This approach allowed inferences about the interaction with room acoustics, enriching the study of musical performances in everyday practice condition where reverberant environments are always present. We tested the approach in a case study using recordings of an excerpt of a clarinet piece of the traditional classical repertoire, played by six professional musicians. A 2D confrontation of the extracted components was proposed. MANOVA tests indicated significant differences ($p<0.05$) when considering the musician as a factor. We examined different *legato*, as well as articulated note transition presenting different performance technique demands.

Keywords: Note attacks description; Music information retrieval; Empirical performance analysis; Spectral modeling.

1 Introduction

Research studies related to the note attack region in musical signals have been conducted since the 1960s. The importance of the attack for the perception of the note was demonstrated in several studies [8, 10, 5]. In [8], for example, it was shown that the musical instrument identification was possible with only 60 ms of the attack, whereas more than twice (150 ms) was necessary using only the note sustain.

Experimental studies on spectral content analysis of musical note attacks have been usually conducted in low reverberant environments, in order to better isolate parameters variations, disregarding the interaction between musician

* This work was supported by CNPq (Brazilian National Council for Scientific and Technological Development).

and performance environment. Adapting the performance to a specific room or concert hall acoustics is part of the musicians routine. In [7], an experiment with professional musicians revealed that they consciously adjust their performing style under different room acoustic conditions. An interview with the performers indicated that each participant consciously adjusted the performing style according to some of the following attributes: tempo, vibrato, harmonics, sound quality, articulation, agogics, and dynamics. Also, studies have focused on the objective investigation of adjustments of performances made by professional musicians under different room acoustic conditions [7, 12, 1, 4]. In [12] for example, computer models of 14 rooms were tested by dynamic binaural synthesis with performers playing physically in an anechoic chamber, while wearing extra aural headphones simulating corresponding virtual acoustic environments. This experiment revealed distinct concepts of adjustment to room acoustical conditions as well as significant individuality with respect to musicians interaction with room acoustics.

Inspired by this, we explored the audio content of monophonic musical note attacks, considering the reverberated harmonics of the previous note as a relevant information. We applied Spectral Modeling Synthesis (SMS) decomposition [13] for estimating the energy of the sinusoidal and residual components, using this information to audio description as proposed by [6]. Here, however, we added a step in the modeling and considered the harmonic content of reverberation independently from the harmonic content of the note being played. This approach suggested inferences about the interaction with room acoustics, enriching the study of musical performances in everyday practice condition where reverberant environments are always present. A 2D confrontation of the estimated components was proposed for visual data analysis. The analysis was conducted for different performance situation of note articulation, such as *legato* and *staccato*, as well as in note transitions demanding different performance techniques. Comparing the efficiency of different modeling techniques goes beyond the scope of this article.

2 Methods

In an overview, we can divide the proposed methodology into three steps: (i) audio segmentation (pre-processing); (ii) spectral modeling and decomposition; (iii) characterization of attacks using the extracted components.

2.1 Segmentation

All the audio processing was done using MATLAB from MathWorks. Figure 1 shows the audio segmentation processing. First note onsets are detected on the RMS signal combining energy and pitch criteria. We used Hamming window with size of 1024 samples (23.2 ms) with hop size of 256 samples (5.8 ms), at the sample rate 44.1 kHz. F0 was estimated by choosing the local maximum of the spectrum that presents the highest harmonic series energy sum. Onset

of consecutive notes with equal pitch were detected by crossing the RMS curve by another RMS with window size of 1 second. Transitions involving notes with different pitches, onsets were detected by variations in the pitch curve above 6%. Techniques for transients detection in musical signals are based on four principles: variations in signal energy, variations in frequency spectrum magnitude, variations in phase spectrum and detection by modeling [3]. Here, variations in frequency spectrum magnitude was chosen to detect the attacks by means of Spectral Flux, defined as the correlation coefficient of the frequency spectrum magnitude between consecutive frames, according to equation 1.

$$F(q) = 1 - \frac{1}{M} \sum_{p=1}^{M} |r(X(p)_q, X(p)_{q-1})| \tag{1}$$

where $F(q)$ is the Spectral Flux value in frame q, M is the window size, r is the correlation coefficient, and X is the frequency spectrum of the current frame.

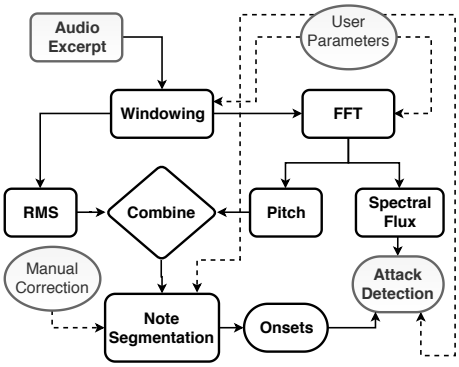


Fig. 1. Diagram of pre-processing.

Spectral Flux value tends to decrease in regions of small variation in the temporal evolution of the frequency spectrum, and to stabilize only in regions of note sustain. The end of attack was estimated as the first instant of Spectral Flux stabilization below a threshold defined as 10% of the average Spectral Flux of the entire note. Attack region was assumed as the region delimited by the note onset and the end of the attack.

2.2 Spectral Modeling and Decomposition

The proposed analysis initially uses the SMS implementation published in [14] to decompose the frequency spectrum of the attack region (Figure 2) into components $H1$, $R1$, $H2$ and $R2$, corresponding to the harmonic and the residual content of the note being played and of the previous note. First, spectral modeling

synthesis $SMS(1)$ models the harmonics energy $H1$ with frequency restriction around the attacked note pitch (Pitch Restriction 1). $R1$ is as the first residue obtained by spectral subtraction. Then, $R1$ is submitted to spectral modeling $SMS(2)$, with frequency restriction around the pitch of the previous note (Pitch Restriction 2), resulting in a second pair of components: $H2$, the energy of the modeled harmonics from reverberation of the previous note and $R2$, the energy of the second residue. Five percent of frequency restriction was adopted. $H1$, $R1$, $H2$ and $R2$ values are expressed in terms of their proportions to the original energy.

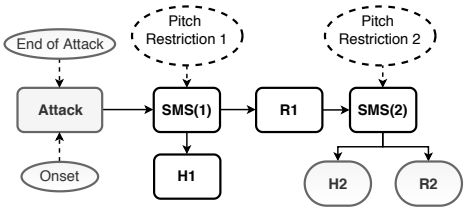


Fig. 2. Diagram of analysis.

2.3 Characterization of Attacks

A 2D confrontation of components $H2$ and $R2$ is proposed for characterizing note attacks, which facilitates the comparison between the energy content due reverberation of the previous note with the harmonic energy of the note being attacked. Higher values of any of these two variables correspond to lower harmonic energy from the current note, since the sum of $H2$ and $R2$ equals to $R1$.

In order to verify whether the proposed representation is capable to describe musicians' reaction to the acoustics of performance ambience, as observed by [12], we carryed out a MANOVA test (with Pillai's trace[1]) to test the effect of musician as a factor on variables $H2$ and $R2$. Variances of the covariance matrix were also used to infer about musicians' consistency. All the statistical analysis was conducted using programming environment R [9].

3 Materials

We tested the approach in a case study using a recording set of selected note transitions of an excerpt of the first movement of Brahms Clarinet Sonata in F minor, op. 120 No. 1 (Figure 3), played by six professional clarinetists. Participants were permanent members of the Philharmonic Orchestra of Minas Gerais and

[1] The Pillais trace is the default test option used in R function MANOVA and considered the most robust statistic for general use, here chosen due to the exploratory characteristics of the research.

advanced clarinet students of the Federal University of Minas Gerais (UFMG). They were all very familiar to the music material used in the experiment. They were asked to play the excerpt, without piano accompaniment, using their own instruments and materials. Recordings were in the same room, with the same equipment equaly configured.

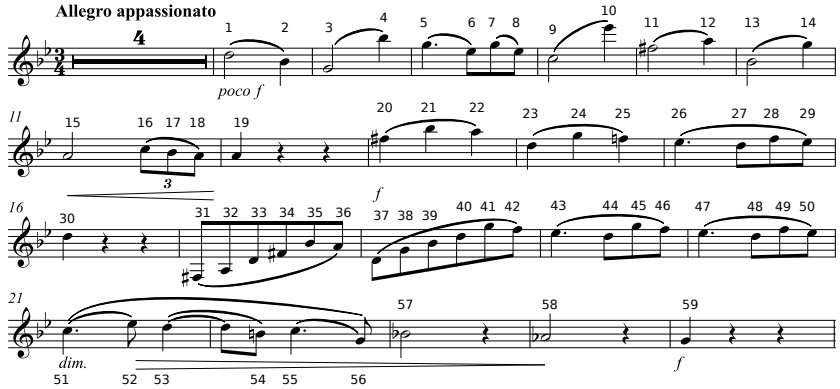


Fig. 3. Excerpt of the first movement of Brahms Clarinet Sonata in F minor, op. 120 No.1, used in the analysis.

4 Results

We selected seven cases of consecutive note transition to examine our representation for characterizing note attacks. The selection was based on different levels of performance technique demands presented by each transition, due to note articulation type and note interval leap. Five of them are *legato* transitions (connected by a slur on the score): notes 3 to 4, 9 to 10, 11 to 12, 13 to 14 and 21 to 22 (Figure 3). Two are articulated (detached) transitions (indicated by a slur break): notes 8 to 9 and 12 to 13.

Figure 4 (a) shows the 2D confrontation for all executions of the transition to note 4. T1 and T2 represent two executions of these transition by the same clarinetist. Figure 4 (b) and (c) show T2 and T1 waveforms, spectrograms and Spectral Flux confirming the distance between both executions observed on the 2D representation, where execution T2 exhibits much lower harmonic content ($H1$) then execution T1. In this figure, waveforms and spectrograms were extracted from the software Sonic Visualizer [2] and Spectral Flux (continuous white line) using the plugin MIR.EDU [11]. The elliptical shapes are data-concentration delimiters. These shapes were estimated as bivariate-normal probability-contour at levels equal to 0.8. Associated solid points in these figures represents the center of the ellipses.

Figure 5 shows the 2D confrontation for the *legato* transitions to notes 10, 12, 14 and 22 and Figure 6 for the articulated transitions to notes 9 and 13.

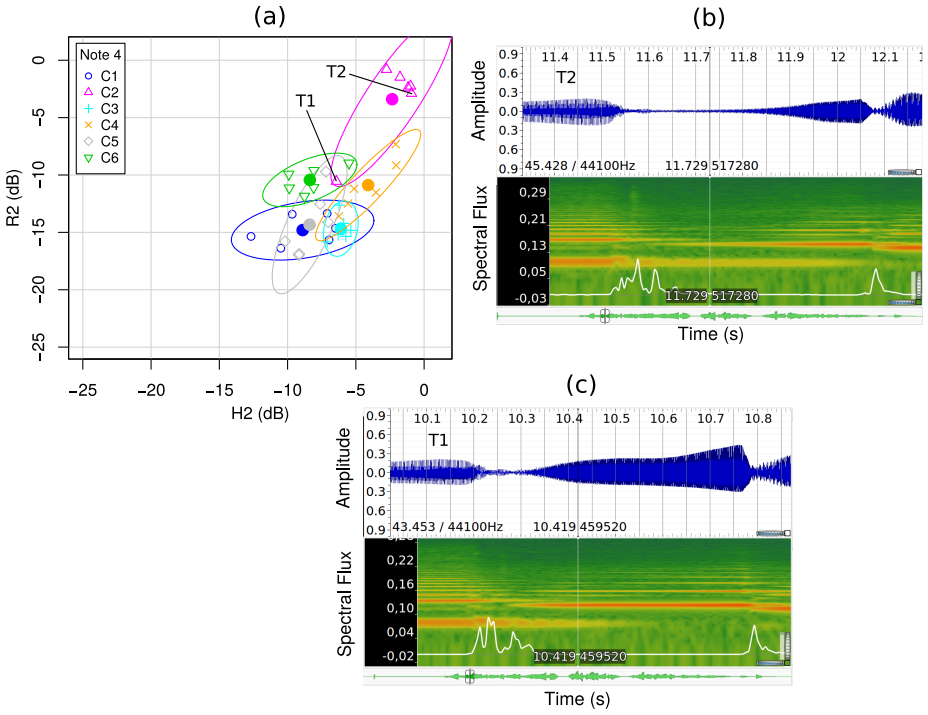


Fig. 4. 2D visualization of the six clarinetists (C1 to C6) executions of transitions to note 4 (panel a) with T1 and T2 indicating two executions by clarinetist C2. The other panels shows the waveform and spectral flux (spectrogram in background) of T2 (panel b) and T1 (panel c).

Table 1 shows the results of MANOVA (p-values) considering the clarinetist (C1 to C6) as a factor for each of the selected notes, and also the variances of $R2$ and $H2$, for each clarinetist by note and, finally, the group pairs with Maximum Contrast.

5 Discussion

Legato note transitions, as to notes 4, 10, 12, 14 and 22, are achieved by maintaining the blowing pressure across the transition, i.e., the reed vibration should not be interrupted neither by any decrease of the blowing pressure nor by touching the reed with the tongue. The quality of *legato* transitions between consecutive notes is quite dependent to the ability of keeping the blowing pressure stability. Moreover, as the width of the interval leap between the notes increases, the execution of the transition may impose additional technical demands that can vary according to the leap width. It may demand precision on synchronizing movements that involve more fingers, sometimes of different hands. Additionally, some leaps may also present extra skills for synchronizing the necessary finger

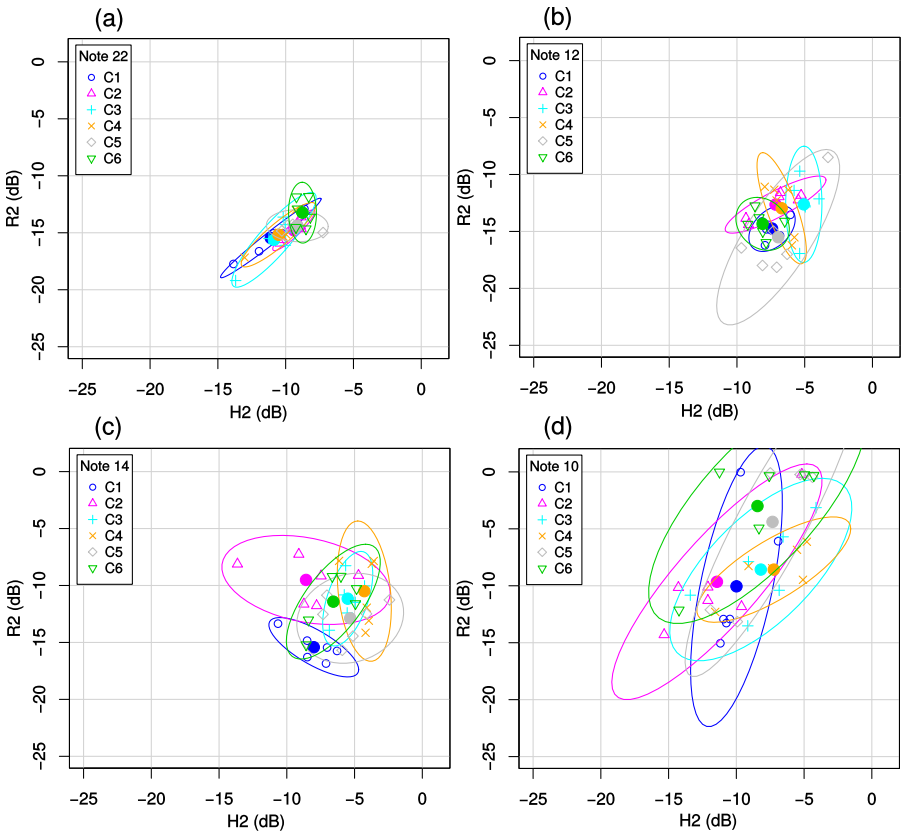


Fig. 5. 2D visualization for selected notes in *legato* condition 10, 12, 14 and 22.

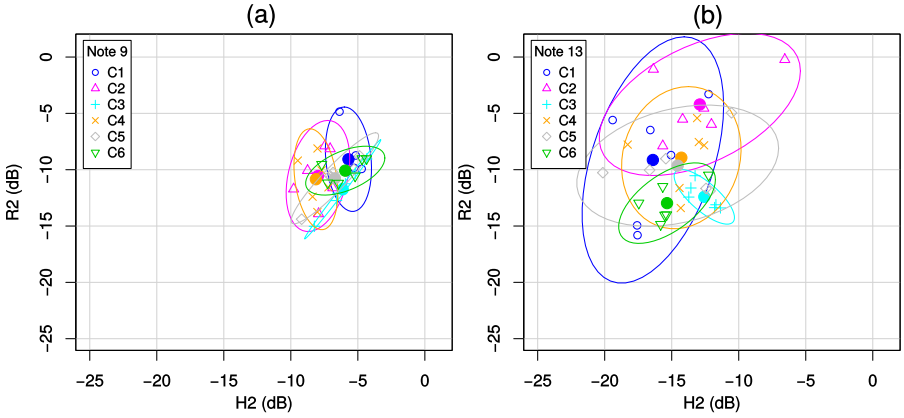


Fig. 6. 2D visualization for articulated transition to notes 9 and 13.

Table 1. Variances of $R2$ and $H2$, for each clarinetist (C1 to C6), by note and group pairs, and also Maximum Contrast obtained from MANOVA testing clarinetists as a factor.

	Variance						Max. contrast	MANOVA
	C1	C2	C3	C4	C5	C6	in dB (pair)	P value
Note 4							4.4 (C4-C5)	$< 0.001^*$
R2	1.5	12.9	1.3	5.2	8.0	1.2	-	-
H2	6.0	4.5	0.4	3.3	1.7	2.6	-	-
Note 9							2.1 (C1-C4)	0.007
R2	4.8	5.3	4.3	4.4	3.6	1.0	-	-
H2	0.7	1.3	1.8	0.6	2.4	1.9	-	-
Note 10							4.8 (C1-C2)	0.015
R2	33.6	23.7	14.2	4.7	40.9	23.6	-	-
H2	2.5	13.3	10.0	7.1	9.2	14.3	-	-
Note 12							2.4 (C3-C5)	0.013
R2	0.9	1.4	5.8	5.3	13.1	1.2	-	-
H2	0.7	3.1	0.4	0.7	4.6	0.9	-	-
Note 13							5.6 (C2-C6)	< 0.001
R2	26.5	8.8	1.3	8.8	6.4	2.8	-	-
H2	6.2	12.5	1.12	4.4	12.6	2.9	-	-
Note 14							4.3 (C4-C5)	< 0.001
R2	1.5	3.4	3.8	8.4	3.5	5.6	-	-
H2	2.4	8.6	0.8	0.9	3.4	2.7	-	-
Note 22							2.3 (C2-C3)	0.027
R2	2.7	0.7	3.8	1.8	0.4	1.6	-	-
H2	3.1	0.5	2.1	1.7	1.2	0.2	-	-

$*$Clarinetist group C2 was excluded from MANOVA in note 4 due to strong rejection in normality test.

movement with the mouth pressure adaptation to the instrument tube vibration change to the next note, especially when the transition involves register change.

The transition to note 22 (Figure 5 a) involves a downward semi tone that requires merely the release of a key to close a hole, driven by a steel spring, meanwhile the transition to note 12 (Figure 5 b) requires lifting up one left hand finger simultaneously with one right hand finger, demanding precision on the synchronization of the movement of both fingers. The graphics show higher harmonic content of the target note 22, hence less reverberation ($H2$) and less noise ($R2$), which may indicate a higher *legato* quality on this transition. We also observe lower variance of both dimensions for note 22, either considering all executions, or the executions within participants. This suggests that players are able to achieve more success in the execution of this transition.

The *legato* to note 14 (Figure 5 c) and to note 4 (Figure 4) involve an upward major sixth and a minor tenth leaps, respectively, crossing from the low to the high register of the instrument. The difficulty in these transitions lies not only on the multiple simultaneous finger action, but also on maintaining the blowing

pressure stable, as well as on synchronizing the mouth pressure coupling with the finger movement. Comparing the executions of both transitions we can observe higher variance of both dimensions for note 4, which might suggest that players have more difficulty to achieve better *legato* quality for that transition. In fact, other than involving wider interval leap, the transition from G4 to B flat 5 (note 4) requires complex actions of simultaneous hole covering and key pressing with fingers of both hands.

The most demanding of these passages is the transition to note 10 (Figure 5 d). Like for note 4, note 10 is reached by an upward minor tenth leap, but crossing from a low note of the high register to the *altissimo* register. This passage imposes additional technical difficulty due to the acoustic instability of the instrument in the altissimo register, where players have always to make use of various alternative fingerings, in order to correct the intonation, which imposes additional demand on synchronizing the mouth pressure coupling with finger movement. It is common for players to fail to execute correctly this passage no matter their degree of expertise, as some of the observed executions.

Comparing the executions of both articulated transitions, notes 9 and 13, Figure 6 (a) and (b) respectively, we can observe that note 9 exhibits higher overall harmonic content of the target note, except for one outlier execution. This might be explained by a longer articulation length imposed by the register change (high to low register) involved in this transition. As stated before, register changes demands more finger actions synchronization and mouth pressure adaptation, that might result in longer articulation time between notes. This might also explain the higher variance of executions of note 13, probably reflecting players attempts to perform the articulation with less sound interruption ("hole") between the notes.

6 Conclusion

In this paper, we explored the idea of describing content of monophonic musical note attacks via Spectral Modeling Synthesis decomposition. More specifically, by comparing the energy proportions of harmonic and residual components independently. For characterizing note attacks, we proposed a confrontation between two components: $H2$ - the energy content due to the harmonic reverberation of the previous note, and $R2$ - the residual component by subtracting the harmonic components of the previous note ($H2$) and the note being attacked ($H1$) from the original signal. The confrontation of these two components in terms of their energy proportions to the original signal, allowed us to infer about three components ($H1$, $H2$ and $R2$) in a 2D visualization. This approach provides more detailed description about the spectral content of the attack region and allowed inferences about the interaction with room acoustics, enriching the study of musical performances in everyday practice condition where reverberant environments are always present. We tested the approach using a recording set of an excerpt of a clarinet piece of the traditional classical repertoire, the first movement of Brahms Clarinet Sonata in F minor, op. 120 No. 1, played by

six professional clarinetists. MANOVA tests for the two variables ($R2$ and $H2$) indicated differences ($p<0.05$) when considering musician as factor for all evaluated note transitions. We examined different *legato*, as well as articulated note transition presenting different performance technique demands. It was observed higher variance of both dimensions for notes with higher performance technique demands. This analysis approach was also able to reveal note attacks that were not successfully executed, as observed for the executions of note 4 by clarinetist C2. Future investigations could be done by attempting to find relations between the components proportions and other factors, such as estimates of acoustic impedance of individual note or variations in the room acoustic. We shall also apply the approach to different wind instruments.

References

1. Bolzinger, S., Warusfel, O., Kahle, E.: A study of the influence of room acoustics on piano performance. Le Journal de Physique IV **4**(C5), C5–617 (1994)
2. Cannam, C., Landone, C., Sandler, M.: Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In: Proceedings of the ACM Multimedia 2010 International Conference. pp. 1467–1468. Firenze, Italy (October 2010)
3. Daudet, L.: A review on techniques for the extraction of transients in musical signals. Computer Music Modeling and Retrieval pp. 219–232 (2006)
4. Fischinger, T., Frieler, K., Louhivuori, J.: Influence of virtual room acoustics on choir singing. Psychomusicology: Music, Mind, and Brain **25**(3), 208 (2015)
5. Grey, J.: Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am **61**(5), 1270–1277 (1977)
6. Herrera, P., Serra, X., Peeters, G.: A proposal for the description of audio in the context of mpeg-7. In: Proc. European Workshop on Content-based Multimedia Indexing. Citeseer (1999)
7. Kato, K., Ueno, K., Kawai, K.: Musicians' adjustment of performance to room acoustics, part iii: Understanding the variations in musical expressions. Journal of the Acoustical Society of America **123**(5), 3610 (2008)
8. Luce, D.: Physical correlates of nonpercussive musical instrument tones. Ph.D. thesis, MIT (1963)
9. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018), `https://www.R-project.org/`
10. Risset, J.: Computer study of trumpet tones. The Journal of the Acoustical Society of America **38**, 912 (1965)
11. Salamon, J., Gómez, E.: Mir. edu: An open-source library for teaching sound and music description. Proceedings of the 15th International Society for Music Information Retrieval (ISMIR), Tapei, Taiwan (2014)
12. Schärer Kalkandjiev, Z., Weinzierl, S.: The influence of room acoustics on solo music performance: An experimental study. Psychomusicology: Music, Mind, and Brain **25**(3), 195 (2015)
13. Serra, X., Smith, J.: Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. Computer Music Journal **14**(4), 12–24 (1990)
14. Zölzer, U.: DAFX: digital audio effects. John Wiley & Sons (2011)

Modeling Human Experts' Identification of Orchestral Blends Using Symbolic Information

Aurélien Antoine, Philippe Depalle, Philippe Macnab-Séguin and Stephen McAdams

Schulich School of Music
McGill University
Montreal, QC, Canada
aurelien.antoine@mcgill.ca

Abstract. Orchestral blend happens when sounds coming from two or more instruments are perceived as a single sonic stream. Several studies have suggested that different musical properties contribute to create such an effect. We developed models to identify orchestral blend effects from symbolic information taken from scores based on calculations related to three musical properties/parameters, namely onset synchrony, pitch harmonicity, and parallelism in pitch and dynamics. In order to assess the performance of the models, we applied them to different orchestral pieces and compared the outputs with human experts' ratings in the Orchestration Analysis and Research Database (orchard.actor-project.org). Using different weights for the three parameters under consideration, the models obtained an average accuracy score of 81%. These preliminary results support the initial developments. Nevertheless, future work aims to investigate the weights of each musical property and to include audio analyses to take into account timbral properties as well.

Keywords: Orchestral Blend, Computer Modeling, Perception, Music Information Retrieval

1 Introduction

Orchestration offers myriad possibilities to combine instruments and to create potential sonic effects that could not necessarily be achievable with a smaller ensemble. This musical practice is a challenging subject that is being investigated from a wide range of research perspectives. The aim of our project is to explore perceptual effects of orchestration in order to understand which are the prominent characteristics. It has been suggested that different perceptual effects can emerge from the combinations of instrumental properties, such as the blending or separation of musical instrument sounds to name but two [5]. The fusion of two or more instruments results in a blend effect in which the sounds can be perceived as belonging to a single musical event [15,14], whereas segregation happens when there is a clear separation between different concurrent sound events [11,9]. This paper focuses on the factors affecting orchestral blend and

introduces our approach for developing computer models capable of identifying groups of instruments involved in blend effects. Previous research has proposed computational models of auditory scene analysis [16] but not specifically for instrument blend effects in a musical context. Our aim was to investigate the extent to which the blend effect could be modeled using only symbolic information in order to formalize the processes involved, which would then be utilized for comparisons between computational outputs and data from perceptual experiments. Such developments could be incorporated into systems designed to perform orchestration analysis from machine-readable musical scores and also in computer-aided orchestration systems.

The remainder of this paper is organized as follows. First, we introduce the different characteristics that contribute to the creation of orchestral blends and define the properties that have been utilized in our models. Section 3 provides technical details of the implementation of the orchestral blend models. Then, we apply the models to a selection of orchestral pieces and compare the outputs with human experts' ratings in order to assess the performance of the models. The results of these analyses are then discussed in Section 5. Finally, the last section presents concluding remarks and suggests different ideas to enrich the current models.

2 Orchestral Blend

As mentioned in the previous section, orchestral blend is the perception of different sounds as being grouped into a single event. This perceptual effect is the result of the fusion of instruments properties, an important aspect in orchestration [2]. Several treatises on orchestration have discussed blending techniques and have suggested methods for combining instruments to create this effect [1,12,13]. Blending techniques can be utilized for enriching the quality of a dominating instrument by adding other instruments or for completely fusing instrument sounds to create a unique mixture [12,13,14]. These treatises usually propose guidance for selecting instruments to conceive orchestral blends. For example, Rimsky-Korsakov suggests that woodwind and string instruments tend to blend well [13].

Blend has also been investigated within perception and cognition research. It has been suggested that different acoustical properties play an important role in grouping musical events into a single stream. Following Gestalt principles, sounds evolving similarly are more likely to be grouped together [3]. For instance, onset synchrony, harmonicity, and similar changes in frequency and amplitude across successive sounds might lead the auditory system to group sound components together [10,6]. Moreover, two or more instruments with a low overall spectral centroid and with close centroids tend to blend better. Characteristics related to the the musicians' performance nuances, the spatial position of the instruments, and the room acoustics also contribute to blend effects [7,8]. In regards to musical properties linked to studies in audio perception and cognition, having instruments playing in a harmonic series, in synchrony, and with perfect paral-

lelism in pitch and dynamics often lead to a blend. An example is shown in Fig. 1, in which oboes 1 and 2 in major thirds and clarinets 1 and 2 in major thirds an octave below the oboes play together in harmonic relations (roughly harmonics 4, 5, 8, 10), in synchrony, and with perfect parallelism in pitch and dynamics. Fig. 2 shows another blend example, in which two flutes, two clarinets, and two bassoons are playing together, with two oboes coming in and out throughout the phrase.

Due to processing only symbolic information, we have disregarded spectral properties and information related to performance, spatial position, and room acoustics for the first phase of this project. Instead, we have decided to process musical properties that can be estimated from a musical score. Therefore, for our models, we have estimated the onset synchrony, harmonicity, and parallelism in pitch and dynamics between instruments, recognizing that timbral aspects derivable from the audio signal often also play a role.

3 Score-based Modeling of Orchestral Blend

This project aims to model human experts' identification of orchestral blends using symbolic information. Thus, the initial stage was to define methods to retrieve the musical information from computer-readable scores before developing models for estimating orchestral blend.

3.1 Retrieving Symbolic Score Information

For computer-readable scores, we chose to work with MusicXML, a format based on the markup language XML used for representing Western musical notation. Several databases offer orchestral pieces encoded in this format. However, they sometimes contain missing or wrong information and also have different templates, which can be due to the software used for their creation. Therefore, we collaborated with OrchPlayMusic (OPM)[1], a company that offers a multichannel audio player for orchestral music and a library of high-quality simulations of orchestral pieces. First, we established a standard procedure to generate MusicXML files with consistent information. We used `MusicXML 3.1`[2] for encoding the symbolic score information. Then, several pieces were selected from the OPM Library, which contains several orchestral excerpts that have been rendered by the OPM team following a set of high-quality simulation techniques, and have been generated as MusicXML files. In order to retrieve the specific musical information required for estimating the characteristics involved in the blend effect, we programmed different functions in `Python 3.7` to process MusicXML files, following the methods further described in the next section.

[1] www.orchplaymusic.com
[2] www.w3.org/2017/12/musicxml31/

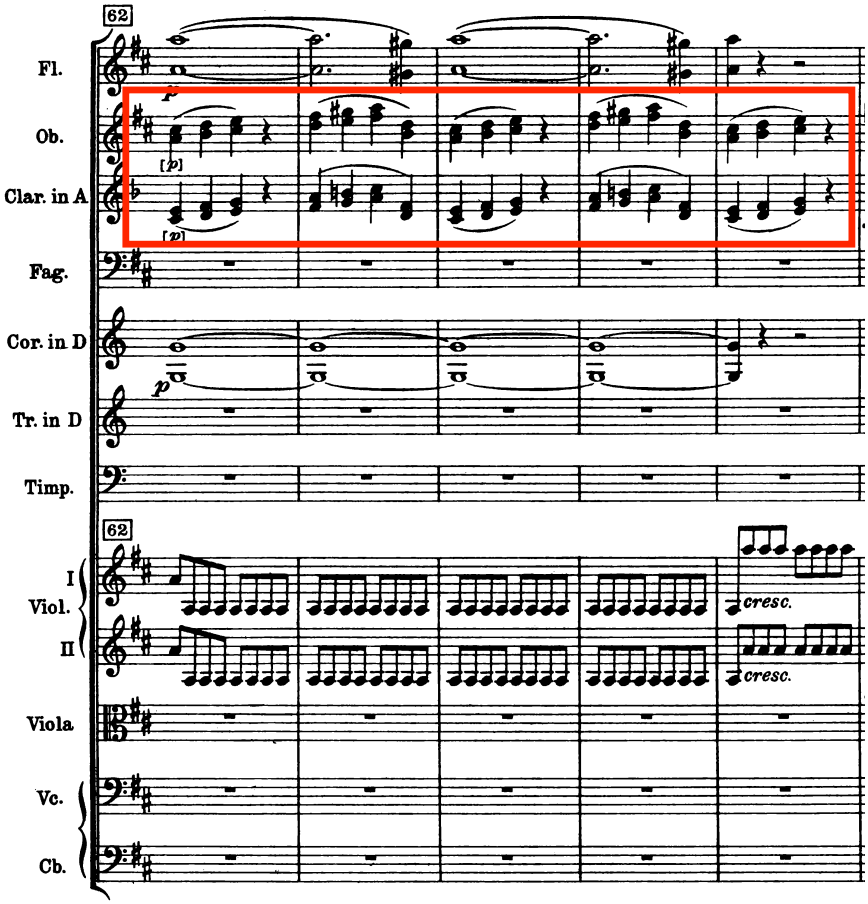


Fig. 1. Example of a blend between two oboes and two clarinets (annotated with a red box) in Mozart, Don Giovanni, mm. 62–66.

3.2 Estimating Orchestral Blend

The process for estimating orchestral blend from computer-readable scores is divided into different steps. First, we had to define the segmentation of the musical pieces. We decided to perform the analysis on a measure-by-measure basis, as orchestral effects most often occur over the course of at least a measure. The rationale is that performing the calculations on a shorter frame (i.e., note by note) would result in a significant increase in the amount of information to compute and compare, and it would also omit aspects related to temporal properties (i.e., parallelism). A longer analysis frame could overlook effects occurring in a single measure. Thus, a measure worth of information appeared to be an appropriate analysis time frame to start with.

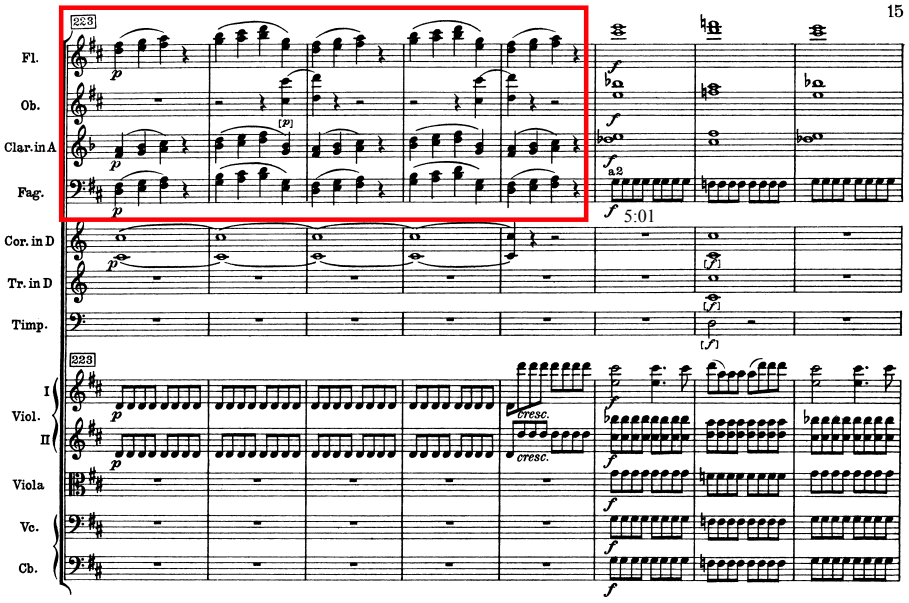


Fig. 2. Example of a blend between two flutes, two oboes, two clarinets, and two bassoons (annotated with a red box) in Mozart, Don Giovanni, mm. 223–227.

The estimations of orchestral blend are computed in three steps: first the onset synchrony, then the pitch harmonicity, and finally the parallelism in pitch and dynamics. Fig. 3 presents a diagram of the different processes for estimating orchestral blend, which are detailed below.

Onset Synchrony. The first step is to list all the onset values for each active instrument in the measure. MusicXML's note duration being represented as divisions per quarter note, it is necessary to convert the symbolic duration into time, defined in milliseconds using tempo values and note types. It is then possible to calculate the onset value for each note. Furthermore, using duration in milliseconds allows us to define a threshold for considering notes being synchronized. We set the default threshold at 30 ms, following suggestions by research on attack time [4]. Thus, notes are considered synchronized if their onset values fall within a 30-ms window. Then, the instruments sharing the most onset values are grouped together. Finally, the synchrony score is calculated with the cardinality of the intersection of the different sets of onset values. The groups of synchronous instruments are then passed to the function for pitch harmonicity calculations. Also, if there is no group of synchronous instruments, the algorithm bypasses the other functions and moves to the next measure.

Pitch Harmonicity. In the second step, this function takes as input the groups of synchronous instruments. For each onset value, the function retrieves the pitch for each active instrument and calculates the interval in semitones between the different pitches. It determines whether they are in a harmonic

series, using the lowest pitch as the root. If the instruments are not all in a harmonic series, the function lists all the different instrument pitches and checks if a tonal chord is involved, following a framework of standard Western musical chords. If no harmonic chord is found, it keeps the largest list of instrument pitches that are in a harmonic series, similar to applying a harmonic template. This step also removes instruments that share onset values but are not in the harmonic series and potentially not involved in the blend. Once the intervals for all the onset values are analyzed, the function returns a harmonicity score for the instruments that are either playing in a harmonic series or forming a tonal chord.

Parallelism in Pitch and Dynamics. The final step looks at the evolution of the pitches and the dynamics over the course of a measure. The function estimates whether the different instruments are playing in parallel. For each instrument, it lists the note sequence by examining if the next note is higher ($+1$), lower (-1), or the same (0) in pitch as the initial note of the measure, which is set to 0. Then, it compares each element of the different lists of note sequences and adds 1 if for each note they all have the same value (i.e. $+1$, -1 or 0) and 0 if at least one is different. The resulting score is then divided by the number of elements in the list, giving us a proportion that is then used for the parallelism score. A similar procedure is applied for the dynamics, where the function examines whether the instrument is playing the notes harder, softer, or at the same dynamic, and then calculates the parallelism.

Output Decision. Once the three properties have been calculated, their corresponding scores are averaged and compared to a defined threshold. If the average score is above the threshold, the group is output as a potential blend. If it is below, the group is ignored and the next group is tested, or the program moves to the next measure if there is no other group of instruments to analyze. A threshold defined as 100 would mean that all the instruments in the group would have perfect synchrony, harmonicity, and parallelism in pitch and dynamics. Lowering the output threshold would allow for more flexibility in the calculations of the musical characteristics and would tolerate deviation from the theoretical rules. This would also account for the different strengths of the blend effect. Furthermore, the properties are set as having the same weights in the calculations. For instance, synchrony is as important as parallelism and harmonicity.

For each measure, the program lists the group(s) of blended instruments with their scores, if a blend has been detected. Nevertheless, orchestral effects can happen over several measures. Thus, we also apply a post-blend analysis function in order to find groups of blended instruments that span consecutive measures. It compares the list of instruments in two neighboring measures and groups them if all the instruments are in both measures. The grouping continues until the instruments are not present in the next measure. The blend is then listed as happening from measure a to measure b, with the names of the instruments involved in the effect. Using the example shown in Fig. 1, the blend occurs from measures 62 to 66, with oboes 1 and 2 and clarinets 1 and 2 playing in every

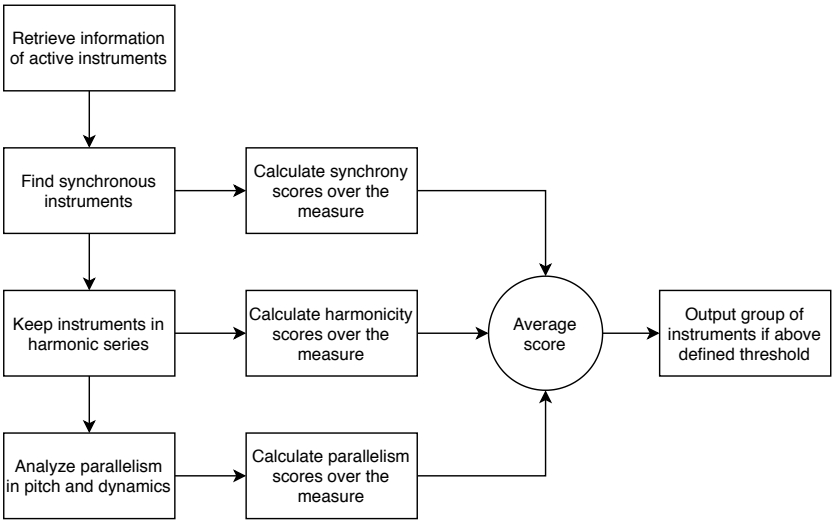


Fig. 3. Synopsis of the orchestral blend estimation algorithm.

measure and returned as blending instruments. Here, the model would return the group of instruments (oboe 1-2 and clarinet 1-2) and specifies that the effect starts at measure 62 and finishes at measure 66.

4 Comparison between Score-Based Models and Human Experts' Identification

In order to test and assess the performance of the score-based modeling of orchestral blend, we decided to apply the models to orchestral excerpts and compare the output with blends that have been identified by human experts. Therefore, we decided to use annotations taken from the Orchestration Analysis and Research Database (Orchard)[3], which contains numerous annotations of orchestral effects derived from the analysis of several orchestral pieces, with the majority spanning the period from 1787 to 1943. However, the selection of pieces to perform the testing was restrained to what is available in both the OPM library and the Orchard database. Furthermore, only parts of some of the orchestral pieces are included in the OPM library. In the end, we utilized the MusicXML files of the following 5 orchestral excerpts:

- Berlioz - Symphonie Fantastique - IV (mm. 1–77)
- Mozart - Don Giovanni - Overture (mm. 1–284)
- Haydn - Symphony 100 - II (mm. 1–70)
- Mussorgsky - Pictures at an Exhibition - II (mm. 57–109)
- Mussorgsky - Pictures at an Exhibition - XIII (mm. 1–22)

Table 1. Blend-detection score using an output decision threshold set at 100 for a blend in Mozart, Don Giovanni, mm. 62-66, shown in Fig. 1.

Measure number	Human experts	Model	Score (ratio)
62	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	1.0
63	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	1.0
64	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	1.0
65	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	1.0
66	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	Oboe 1, Oboe 2, Clarinet 1, Clarinet 2	1.0

Table 2. Blend-detection score using an output decision threshold set at 80 for a blend in Mozart, Don Giovanni, mm. 223-227, shown in Fig. 2.

Measure number	Human experts	Model	Score (ratio)
223	Flute 1, Flute 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	Flute 1, Flute 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	1.0
224	Flute 1, Flute 2, Oboe 1, Oboe 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	Flute 1, Flute 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	0.75
225	Flute 1, Flute 2, Oboe 1, Oboe 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	Flute 1, Flute 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	0.75
226	Flute 1, Flute 2, Oboe 1, Oboe 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	Flute 1, Flute 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	0.75
227	Flute 1, Flute 2, Oboe 1, Oboe 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	Flute 1, Flute 2, Clarinet 1, Clarinet 2, Bassoon 1, Bassoon 2	0.75

The comparison process was designed to evaluate if the models were retrieving the same groups of instruments as the ones labeled as part of a blend by the human experts. For example, Table 1 details the results of the comparison for a blend identified in the Overture to Mozart's Don Giovanni (shown in Fig. 1). Here, using an output decision threshold set at 100, the models have output the same group of instruments (oboe 1-2 and clarinet 1-2) as the human experts for the five measures, obtaining a score of 4/4. Table 2 details the results for the blend shown in Fig. 2, also taken from Don Giovanni. Here, the output decision threshold was set at 80. Note that the models have output the same instruments as the ones labeled by the experts for measure 223 (i.e., flute 1-2, clarinet 1-2, and bassoon 1-2), resulting in a score of 6/6. However, for the measures 224 to

[3] `https://orchard.actor-project.org`

Table 3. Summary of the average ratio scores (in %) and number of blends for each and across all orchestral pieces.

	Average ratio scores (in %)			
Musical pieces	**Output decision threshold = 100**	**Output decision threshold = 80**	**Output decision threshold = 60**	**Number of blends**
Berlioz - Symphonie Fantastique - IV	49.57	81.53	77.81	8
Mozart - Don Giovanni	41.29	91.04	91.34	17
Haydn - Symphony 100 - II	71.60	88.93	89.88	7
Mussorgsky - Pictures at an Exhibition - II	70.48	70.64	70.64	17
Mussorgsky - Pictures at an Exhibition - XIII	35.91	77.75	76.20	7
Across all pieces	53.77	81.98	81.17	56

227, the models missed the oboe 1-2, thus, obtaining a proportional score of 0.75 (6/8).

A similar comparison process was applied to the whole testing set. We ran the models using output decision thresholds set at 100, 80, and 60, in order to investigate if more flexibility in the calculations would improve the accuracy or not. Table 3 proposes a summary of the average proportion scores, expressed in %. The performances for the three different output thresholds are listed for each orchestral excerpt and across all the pieces, along with the number of blends for each piece. Note that the best average score was obtained using an output decision threshold set at 80, which resulted in an accuracy of 81.98% for the 56 blends labeled across all the orchestral pieces.

5 Discussions

From the results detailed in the previous section, it is clear that using only the standard rules, represented by defining the output decision threshold at 100, is not sufficient. We note a significant improvement between the output decision thresholds set at 100 and at 80, while the difference between 80 and 60 is small. As detailed in Table 3, the average accuracy score across all the pieces using the threshold at 100 was 53.77%, whereas it was over 81% with thresholds set at 80 and 60. For the symbolic information from Mozart - Don Giovanni and Haydn - Symphony 100, movement II, the models performed better with the lowest threshold. In regards to the two movements of Mussorgsky - Pictures at an Exhibition, the models did not improve when the output decision threshold was set at a lower value. The performance even decreases when setting the threshold at 60 for Berlioz - Symphonie Fantastique, movement IV, suggesting that more flexibility in the rules may have created confusion in the models.

Some limitations of this initial implementation have emerged from the comparison process. For instance, in the blend shown in Fig. 2, the models missed

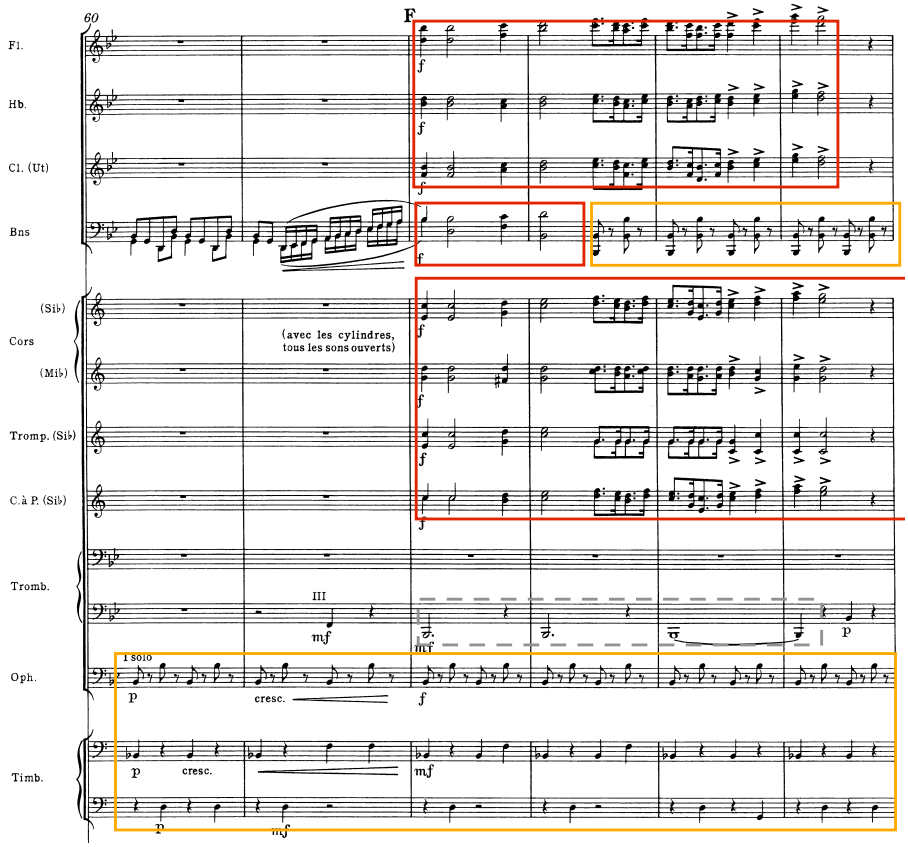


Fig. 4. Example of a blend between a tuba, two timpani, and four bassoons later in the phrase (annotated with an orange box), mm. 60–65, and a blend between two flutes, two oboes, two clarinets, four bassoons, four horns, two trumpets, and two cornets (annotated with a red box), mm. 62–65, in Berlioz, Symphonie Fantastique IV

the oboe 1-2 in the measures 224 to 227. This is due to them playing one note in each measure, and thus, having a low score for onset synchrony (1 common onset value out of 4 with all the other instruments) as well as for the parallelism properties. The current implementation is not able to retrieve instruments involved in a blend and playing sporadically compared to the rest of the blended instruments. Another limitation is illustrated with Fig. 4. Here, the four bassoons switch from one group of blend (annotated with a red box) to an other (annotated with an orange box) in the middle of the measure 63. The models have grouped the bassoons with the flutes, oboes, clarinets, horns, trumpets, and cornets instead of with the tuba and timpani for the measure 63 and 64. Due to performing the analysis on a measure-by-measure basis, the models cannot notice a change of blended group if it occurs within a measure. Furthermore, given that the pitch harmonicity function is based either on harmonic series of

semitone intervals or on a succession of tonal chords, if instruments play notes that are do not follow this framework, they would be discarded. This could be another reason that the models did not output all of the instruments involved in an orchestral blend.

Although our models have accurately output almost 82% of blended instruments on average across 56 blends, the results detailed in Section 4 indicate that processing only symbolic information is not enough to thoroughly model the orchestral blend effect.

6 Conclusion and Future Directions

In this paper, we have presented our approach for modeling human experts' identification of orchestral blends using symbolic information from computer-readable scores. Our partnership with OrchPlayMusic has allowed us to get standard, precise, and consistent MusicXML files of orchestral pieces from which to process symbolic score information. We based our models on the evaluation of three musical characteristics suggested by previous research on orchestral blend: onset synchrony, pitch harmonicity, and parallelism in pitch and dynamics, as described in Section 3. In order to evaluate the performance of the models, we decided to compare their outputs with blends labeled by human experts. Therefore, we selected and processed five orchestral pieces that had been previously analyzed by musical experts and also generated as MusicXML files by the OPM team. As detailed in Section 4, the models achieved an average accuracy score of 81% across all the pieces.

Preliminary results support the initial developments and suggest that estimations based on symbolic information can account for a significant part in modeling orchestral blends. However, further investigation is required to overcome the current limitations discussed in Section 5. For instance, tuning the weights of the different calculations could be an aspect to consider, as perhaps onset synchrony is a more prominent characteristic than pitch harmonicity. The use of supervised machine learning techniques combined with a large set of labeled blend examples could potentially aid in addressing this question. Moreover, spectral, temporal, and spectrotemporal audio properties also contribute to the blend effect, as mentioned in Section 2. Therefore, future work also aims to include audio analyses to take into account timbral characteristics as well.

References

1. Berlioz, H.: Grand traité d'instrumentation et d'orchestration modernes. Henry Lemoine, Paris, France (1844)
2. Blatter, A.: Instrumentation and orchestration. Schirmer Books, New York, NY, 2nd edn. (1997)
3. Bregman, A.S.: Auditory scene analysis: The perceptual organization of sound. MIT press, Cambridge, MA (1990)
4. Bregman, A.S., Pinker, S.: Auditory streaming and the building of timbre. Canadian Journal of Psychology/Revue canadienne de psychologie 32(1), 19 (1978)

5. Goodchild, M., McAdams, S.: Perceptual processes in orchestration. In: Dolan, E.I., Rehding, A. (eds.) The Oxford Handbook of Timbre. Oxford University Press, New York, NY (2018)
6. Kendall, R.A., Carterette, E.C.: Identification and blend of timbres as a basis for orchestration. Contemporary Music Review 9(1-2), 51–67 (1993), https://doi.org/10.1080/07494469300640341
7. Lembke, S.A.: When timbre blends musically: Perception and acoustics underlying orchestration and performance. Ph.D. thesis, McGill University, Montreal, QC, Canada (2014)
8. Lembke, S.A., Parker, K., Narmour, E., McAdams, S.: Acoustical correlates of perceptual blend in timbre dyads and triads. Musicae Scientiae pp. 1–25 (2017), https://doi.org/10.1177/1029864917731806
9. McAadams, S.: Musical timbre perception. In: Deutsch, D. (ed.) The psychology of music, pp. 35–67. Academic Press, San Diego, CA, 3rd edn. (2013)
10. McAdams, S.: The auditory image: A metaphor for musical and psychological research on auditory organization. In: Crozier, W.R., Chapman, A.J. (eds.) Cognitive Processes in the Perception of the Art, pp. 289–323. Elsevier, North-Holland, Amsterdam (1984)
11. McAdams, S., Bregman, A.S.: Hearing musical streams. Computer Music Journal 3(4), 26–43 (1979)
12. Piston, W.: Orchestration. WW Norton, New York, NY (1955)
13. Rimsky-Korsakov, N.: Principles of orchestration. Dover Publications, New York, NY, 1st edn. (1964)
14. Sandell, G.J.: Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. Music Perception: An Interdisciplinary Journal 13(2), 209–246 (1995), https://doi.org/10.2307/40285694
15. Sandell, G.J.: Concurrent timbres in orchestration: a perceptual study of factors determining "blend". Ph.D. thesis, Northwestern University, Evanston, Illinois, United States of America (1991)
16. Wang, D., Brown, G.J. (eds.): Computational auditory scene analysis: Principles, algorithms and applications. Wiley-IEEE Press, New York, NY (2006)

The Effect of Auditory Pulse Clarity on Sensorimotor Synchronization

Prithvi Kantan, Rareș Ștefan Alecu and Sofia Dahl

Aalborg University Copenhagen
[pkanta18,ralecu18]@student.aau.dk; sof@create.aau.dk

Abstract. This study investigates the relationship between auditory pulse clarity and sensorimotor synchronization performance, along with the influence of musical training. 27 participants walked in place to looped drum samples with varying degrees of pulse clarity, which were generated by adding artificial reverberation and measured through fluctuation spectrum peakiness. Experimental results showed that reducing auditory pulse clarity led to significantly higher means and standard deviations in asynchrony across groups, affecting non-musicians more than musicians. Subjective ratings of required concentration also increased with decreasing pulse clarity. These findings point to the importance of clear and distinct pulses to timing performance in synchronization tasks such as music and dance.

Keywords: pulse clarity, sensorimotor synchronization, rhythm, movement, perception, musical training, timing

1 Introduction

Sensorimotor synchronization (SMS) [1] is a form of referential behavior in which an action is coordinated with a predictable external event, the referent, where both are usually periodic. Examples of SMS are dance (where movements are synchronized with both music and the movements of other dancers), music performance and marching. Although some of the many studies of SMS have involved activities with locomotion and limb movement [2] [3] , the vast majority of studies primarily use finger tapping [1] [4].

SMS studies tend to use auditory stimuli consisting of brief tones or clicks. Such stimuli generally have sharp temporal profiles, which along with low noise in testing environments are likely to exhibit prominent and effortlessly perceptible periodicities or *pulses*. Real-life SMS referents, however, such as music performed in a reverberant hall, often have less pulse salience due to time-smearing and masking effects. Dynamics processing and speaker distortions can further undermine the strength of rhythmic pulsations. In extreme situations, the pulse may no longer even be readily perceptible, subject to individual perceptual ability and musical training or experience.

Pulse Localization and Alignment. Ecological stimuli may exhibit varied temporal envelopes with distinct sub-band spectral evolution. The *perceptual center* (P-center) of a sound is understood as the specific moment of perceived occurrence [5]. Synchronization then involves aligning P-centers, which studies [5] [6] have shown to depend upon envelope characteristics. The P-center seems to be located between the perceptual onset and the energy peak of a sound. For impulsive sounds this is close to a single location, while tones with gradual onsets tend to show a range of equally 'correct' sounding locations [6].

A ubiquitous finding from SMS studies is that taps tend to precede referent tones, a phenomenon called *negative mean asynchrony or NMA* [4]. Trained musicians exhibit both smaller NMA [7] and a smaller standard deviation in asynchrony than non-musicians [8]. Danielson et. al. [5] found that in general, sounds with slower attacks and longer durations had later P-centers with greater variability in their exact location. In a synchronization task with musical and quasi-musical stimuli, NMA with respect to the physical onsets was small to non-existent, but significant with respect to the P-center. This aligned with the hypothesis of Vos et. al. [9] that participants use P-centers, rather than the physical onset of the tone, as the target for SMS tasks.

Full Body Synchronization. There is evidence that pulse salience affects corporeal movement characteristics during SMS tasks [2, 3]. Burger et. al. [3] found that music with greater pulse clarity elicited greater temporal and spatial regularity in dance. Van Dyck et. al. [10] found that when the bass drum was made louder in the music mix, dancers increased their motor activity and entrained better to the beat. In addition to exploring internal SMS mechanisms [11] [12], neuroscience studies show that the extent to which cortical or subcortical motor activations are coupled with the auditory cortex depends on beat salience and music training [13].

As stable pulse perception underlies all SMS tasks, lowering the perceived pulse salience of a referent is likely to have a detrimental impact on SMS performance. A high-level musical measure conveying how easily listeners can perceive the underlying rhythmic or metric pulsation is *Pulse Clarity (PC)* [14]. In this study we used computational methods for pulse clarity estimation [14], to design suitable stimuli for the experimental investigation of relations between pulse clarity and SMS task performance. In-phase walking in place was chosen for greater ecological validity with regard to general bodily movement.

2 Stimulus Design

The creation of suitable auditory referents necessitated 1) The design of rhythmic stimuli spanning the entire range of perceptual pulse salience by systematic manipulation of a single base stimulus. 2) The objective assessment of these stimuli by PC measures.

We made drastic changes to perceived pulse salience by simply altering the decay time of a digital reverberation (reverb) effect with a flat frequency decay, applied to a looped snare drum sample (EZDrummer 2 VST instrument).

The reverb plugin used was the 64 bit version of WAVES TrueVerb, with early reflections and high frequency roll-off disabled.

To determine the range of reverb decay times that fit the required perceptual range, we blindly adjusted decay time to yield pulses that were subjectively *'Very Clear', 'Moderately Clear', 'Moderately Unclear' and 'Very Unclear (but perceptible)'*. We then analyzed these four preliminary sample points of the perceptual range using the MIRToolbox and 1) Entropy of *Onset Detection Function (ODF)* Autocorrelation [14]; 2) Max ODF Autocorrelation; 3) Min ODF Autocorrelation; and 4) Peakiness of Fluctuation Spectrum [2]. The last approach estimates PC by the relative Shannon entropy of the fluctuation spectrum [15], in terms of peak magnitude, regularity of spacing and noise between peaks. The calculated values can be seen in Table 1.

Table 1: Comparison between different methods for PC computation [14] [16]

Sr. No.	Salience	Fluct. Spectrum Peak	EntropyAutocor	MinAutocor	MaxAutocor
1	Very Clear	293961.77	0.5066	0.398	0.9899
2	Moderately Clear	146677.66	0.6314	0.335	0.9749
3	Moderately Unclear	24224.66	0.721	0.3891	0.7672
4	Very Unclear (but Perceptible)	17082.21	0.7439	0.3342	0.5134

Similar to what was reported in [2], the fluctuation spectrum of the perceptually clearer stimuli exhibited peaks with markedly higher magnitude at the beat frequency, and less inter-peak noise than the unclear stimuli (see Figure 1). PC calculations based on Entropy and Min/Max values of ODF autocorrelation, on the other hand seemed to show less agreement with our perceived impressions. The difference could be credited to the presence of perceptual modeling in the fluctuation spectrum calculation [16], absent in the ODF-based methods. Hence, we used the fluctuation spectrum peakiness to model the perceptual range.

From the total range of reverb times, we empirically found that nine total stimuli would sample the perceptual range with enough inter-stimulus difference to minimize redundancy. To determine the necessary fluctuation peak magnitudes, we fitted a 3^{rd} order polynomial curve to the four previously determined values and designed stimuli S1 - S9 to match nine equi-spaced curve values in the same range, in decreasing order of PC. Tempo was centered around 120 BPM, close to the preferred human movement tempo [17], but varied by ±1 BPM between stimuli to prevent short-term training effects [18]. The onset peak amplitude was kept constant across stimuli.

3 Experiment

To test our hypothesis that lower PC would lead to smaller NMA due to perceptual onset masking, and greater variability in SMS performance due to reduced beat salience, we performed the following within-participant experiment:

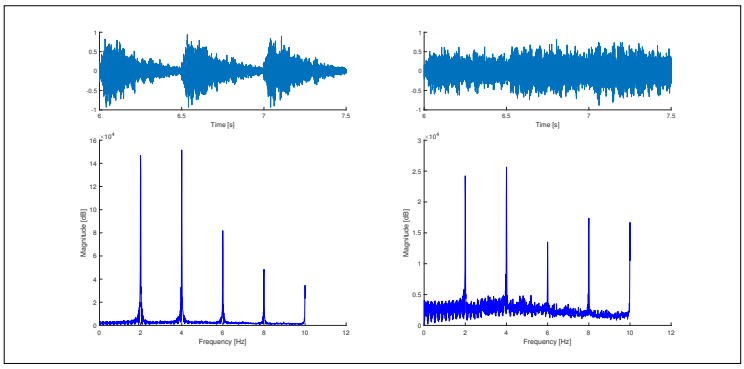


Fig. 1: Pulse clarity measurement using fluctuation spectra. For the signal to the right, the spectrum peaks have lower overall magnitude, and greater noise in between peaks, implying lower pulse clarity. This is also evident from the time domain waveform.

Participants. A convenience sample of 29 participants (6 women, 21-35 years, $MeanAge = 26$), mainly students at Aalborg University, volunteered in return for a film voucher. Participants were briefed on the length of the experiment (9 stimuli x 50 seconds) and that they could withdraw at any time without losing their remuneration.

Experimental Setup and Procedure. Participants were tested individually in a quiet, medium-sized room on campus. The stimuli was played via a set of Focusrite Studio Headphones, while recordings of the activity were captured with a Focusrite CM25 large-diaphragm cardioid condenser microphone. The audio was digitized to a 44.1 KHz/24-bit WAV format using a Focusrite Scarlett Solo Studio audio interface.

After obtaining the participants' informed consent, we asked them to complete an online musical background questionnaire to determine their Ollen Musical Sophistication Index (OMSI) [19]. The OMSI reflects the probability that a music expert would categorize a respondent as *"more musically sophisticated"*, with regard to musical knowledge, skill, and composition ability.

Subsequently, participants were instructed to assume a standing position in front of the microphone and walk in place, stepping in exact synchronization with each of the stimuli, which were presented in random order with brief pauses in between. After each trial, participants were asked to rate on a scale of 1-10 the amount of active concentration required to maintain synchronization with the stimulus. This procedure was repeated for all 9 stimuli.

Data Analysis. Recordings of two participants were discarded in entirety due to poor signal quality, yielding 27×9 trials $= 243$ recordings for analysis. The participants were segregated on the basis of their OMSIs into 3 *MSoph* groups G1 (OMSI $<$ 100, 10 participants), G2 (OMSI 100 - 500, 10 participants), and

G3 (OMSI 500 - 1000, 7 participants), so as to study the effect of musical training on task performance.

The extraction of footstep timestamps was carried out in MATLAB using an onset detection algorithm, and the first 10 seconds of each recording were discarded. From the timestamps, *mean asynchrony (MA)*, *standard deviation - asynchrony (STD-A)* and *inter-tap interval coefficient of variation (ITI-CoV)* were compared across stimuli (S1 - S9) using mixed-design ANOVAs, with 'Stimulus' as the within-subject factor and 'MSoph Group' as the between-subjects factor. A Friedman Test was conducted on the participants' concentration ratings. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. All statistical analysis was done in SPSS 25.0 (IBM Corp).

4 Results

Figure 2 shows the average MA across participants (left) and STD-A (right) for each of Stimulus 1 to 9 (S1 to S9, decreasing PC) The effects of the independent variables on each dependent variable are considered in turn.

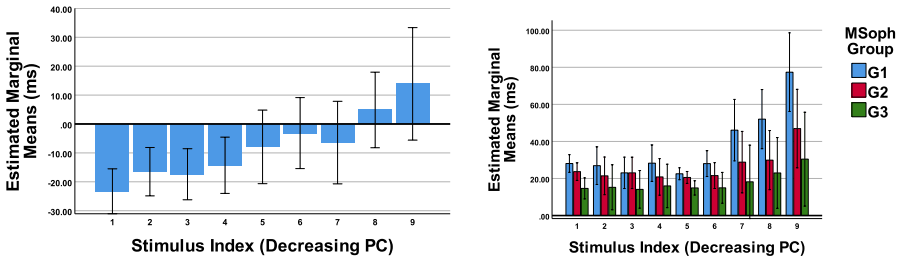


Fig. 2: Average MA across all participants *(left)* and group wise STD-A *(right)* for Stimuli 1 (maximum PC) through 9 (minimum PC). Vertical axis ticks are in ms and 95% confidence intervals are shown in error bars.

We tested the hypothesis that lower PC would lead to smaller NMA in a 9 Stimuli × 3 MSoph Groups mixed-design ANOVA. Results showed a significant main effect of stimulus ($F(2,24) = 7.351$, $p = <.0001$, $\eta_p^2 = 0.776$) and no significant interaction between stimulus and MSoph Group ($F(2,24) = 1.379$, n.s., $\eta_p^2 = 0.393$). A Tukey post-hoc test revealed that MA was statistically significantly more negative for stimulus S1(-23.3 ± 19.6 ms) as compared to S7(-6.4 ± 36.13 ms, $p = 0.028$), S8(4.8 ± 33.1 ms, $p <.0001$), and S9(13.9 ± 49.1 ms, $p< .0001$, see Figure 2(a)). There were no significant differences in pairwise comparisons between MSoph groups.

Another mixed-design ANOVA tested the hypothesis that STD-A would increase with decreasing PC. We found a main effect of stimulus ($F(2,24) = 5.628$, $p =.001$, $\eta_p^2 = 0.726$) with no significant interaction between stimulus and MSoph group. Post-hoc pairwise comparisons showed that STD-A was significantly less

for S1(23.0 ± 8.8 ms) than for S9(53.9 ± 36.8 ms, p =.002), with a clear positive trend from S7 onward. On the basis of MSoph, significant and nearly-significant differences exist between G1 (lowest OMSI group) and G3 (highest OMSI group) (p = 0.02), and G1 and G2 (p =.069) respectively.

The mixed ANOVA for ITI-CoV revealed a significant main effect of stimulus ($F(2,24)$ = 3.964, p =.008, η_p^2 = 0.651), although pairwise post-hoc tests showed non-significant differences between stimuli and MSoph groups. For the stimulus-wise subjective concentration ratings, a Friedman Test found significant differences among stimuli (χ^2 = 144.12, p <.001), and Dunn-Bonferroni-based post-hoc comparisons showed significant differences between multiple pairs of stimuli and non-significant differences between MSoph groups, with a general increasing trend from S1 to S9.

5 Discussion

The purpose of this study was to investigate the relationship between auditory pulse clarity and SMS performance (as measured by asynchrony and normalized variability) as well as the impact of music training on this relationship. Results from the mixed ANOVAS support our hypotheses that lower PC would lead to smaller NMA, and greater STD-A. For high PC stimuli, MA and ITI-CoV parameters agree well with finger-tapping literature [4] (see Figure 2). Measured MA for all MSoph groups showed an increasing trend, with G2 and G3 having more positive mean values. This could be attributed to the masking of *true* perceptual onsets by previous reverb tails, and increased stimulus duration [5].

The clear trend of higher subjective concentration ratings with decreasing pulse clarity indicates that participants attended more closely to less clear stimuli to deduce their underlying pulsations, and maintain their level of synchronization performance. The ratings corroborate the good correspondence we found between fluctuation spectrum peakiness and perceived pulse salience.

The increase of STD-A with decreasing PC was evident beyond S7. An explanation is that the diminished extent of amplitude fluctuation of these stimuli during transients implied a smaller attack slope within the auditory temporal integration window, leading to a similar temporal P-center spread to those observed for slow-attack sounds by Danielsen et. al (2019) [5]. This performance degradation, in particular the greater deterioration for nonmusicians resembles the effect of increasing rhythmic complexity observed by Chen et. al. (2008) [20]. Thus, PC reduction by acoustic degradation may have similar effects on SMS as lowered beat salience due to greater rhythmic complexity.

Our hypothesis that ITI CoV would increase with decreasing PC was not supported by data. In particular for G1, ITI CoV did not increase proportionally with STD-A, as would be expected given the constant stimulus period. The only explanation for these conflicting data is an incorrect overall stepping tempo due to a complete inability to perceive and follow the pulse, particularly for S9.

Overall, these findings indicate that the clarity of the periodic referent has a considerable influence on SMS performance, which would have direct implica-

tions for music and dance performance. The type of degradation present in our stimuli bears resemblance to what might appear in real environments where music or dance activities are performed. For lower PC, beat entrainment not only consumes more cognitive resources, but is also less accurate and stable to a perceptible extent (mean STD-A of 54 ms for S9 v/s 23 ms for S1), highlighting the importance of clear pulse audibility for timing during performance. Interestingly, performance appeared to remain fairly consistent until a certain 'threshold' was crossed, around S7 ('Moderately Unclear') (see Figure 2), implying a certain sensory 'robustness' to referent degradation.

Limitations of the study include the static modality of PC manipulation, short length of the trials and relatively small number of participants, particularly with extensive musical training. Pulse degradations in real-life situations may be time-varying due to the changing spectral content of referents, and unpredictably varying masking effects. Interpersonal entrainment and visual cues during group performance are also important factors. Further studies can address whether the different types of pulse degradation similarly impact SMS performance, and whether these can be accurately modeled by fluctuation spectrum measurements.

6 Conclusion

The present study concluded that reducing auditory pulse clarity influences sensorimotor synchronization performance, with increases in mean and standard deviations of asynchrony values, and ratings of required concentration. The results showed that pulse degradation affected the performance of musically untrained participants more than trained participants. These results have direct relevance to timing performance in dance and music, although further studies must be conducted on a larger sample, exploring other ecological pulse degradation methods to explore their true implications for real SMS contexts.

Acknowledgements. We thank the participants in our experiment. Authors PK and RSA were mainly responsible for the experiment, data analysis and writing of the manuscript. Author SD supervised the project, and assisted in writing. SD's contribution is partially funded by NordForsk's Nordic University Hub Nordic Sound and Music Computing Network NordicSMC, project number 86892.

References

1. Repp, B. H.: Sensorimotor Synchronization: A Review of the Tapping Literature. Psychonomic Bulletin & Review 12(6), 969–992 (2005)
2. Burger, B., Thompson, M. R., Luck, G., Saarikallio, S., Toiviainen, P.: Influences of Rhythm- and Timbre-related Musical Features on Characteristics of Music-induced Movement. Frontiers in Psychology 4, 183 (2013)

3. Burger, B., Thompson, M. R., Luck, G., Saarikallio, S., Toiviainen, P.: Music Moves Us: Beat-related Musical Features Influence Regularity of Music-induced Movement. In: The 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, pp. 183–187 (2012)
4. Repp, B. H., Su, Y. H.: Sensorimotor Synchronization: A Review of Recent Research (2006–2012). Psychonomic Bulletin & Review 20(3), 403–452 (2013)
5. Danielsen, A., Nymoen, K., Anderson, E., Câmara, G. S., Langerød, M. T., Thompson, M. R., London, J.: Where Is the Beat in That Note? Effects of Attack, Duration, and Frequency on the Perceived Timing of Musical and Quasi-Musical Sounds. Journal of Experimental Psychology: Human Perception and Performance (2019)
6. Gordon, J. W.: The Perceptual Attack Time of Musical Tones. The Journal of the Acoustical Society of America, 82(1), 88—105 (1987)
7. Krause, V., Pollok, B., Schnitzler, A.: Perception in Action: The Impact of Sensory Information on Sensorimotor Synchronization in Musicians and Non-Musicians. Acta Psychologica 133(1), pp. 28-–37 (2010)
8. Repp, B. H.: Sensorimotor Synchronization and Perception of Timing: Effects of Music Training and Task Experience, Human Movement Science 29(2), pp. 200-–213 (2010)
9. Vos, P. G., Mates, J., van Kruysbergen, N. W.: The Perceptual Centre of a Stimulus as the Cue for Synchronization to a Metronome: Evidence from Asynchronies. The Quarterly Journal of Experimental Psychology Section A 48(4), 1024—1040 (1995)
10. Van Dyck, E., Moelants, D., Demey, M., Deweppe, A., Coussement, P., Leman, M.: The Impact of the Bass Drum on Human Dance Movement, Music Perception: An Interdisciplinary Journal 30(4), 349–359 (2012)
11. Large, E. W.: On Synchronizing Movements to Music, Human Movement Science 19(4), 527–566 (2000)
12. Fujioka, T., Trainor, L. J., Large, E. W., Ross, B.: Internalized Timing of Isochronous Sounds Is Represented in Neuromagnetic Beta Oscillations, Journal of Neuroscience 32(5), 1791–1802 (2012)
13. Chen, J. L., Penhune, V. B., Zatorre, R. J.: The Role of Auditory and Premotor Cortex in Sensorimotor Transformations. Annals of the New York Academy of Sciences 1169(1), 15–34 (2009)
14. Lartillot, O., Eerola, T., Toiviainen, P., Fornari, J.: Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimization. In: ISMIR, pp. 521–526 (2008)
15. Pampalk, E., Rauber, A., Merkl, D.: Content-based Organization and Visualization of Music Archives. In: Proceedings of the 10$^{\text{th}}$ ACM International Conference on Multimedia, pp. 570–579 (2002)
16. MIR Toolbox, University of Jyväskylä, `https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`
17. Fraisse, P.: Rhythm and Tempo. The Psychology of Music 1, 149–180 (1982)
18. Madison, G., Karampela, O., Ullén, F., Holm, L.: Effects of Practice on Variability in an Isochronous Serial Interval Production Task: Asymptotical Levels of Tapping Variability After Training Are Similar to Those of Musicians, Acta Psychologica 143(1), 119–128 (2013)
19. Ollen, J. E.: A Criterion-related Validity Test of Selected Indicators of Musical Sophistication Using Expert Ratings. Doctoral Dissertation, The Ohio State University (2006)
20. Chen, J., Penhune, V., Zatorre, R.: Moving on Time: Brain Network for Auditory–Motor Synchronization is Modulated by Rhythm Complexity and Musical Training, Journal of Cognitive Neuroscience 20(2), 226–239 (2008)

The MUST Set and Toolbox

Ana Clemente[1*], Manel Vila-Vidal[2], Marcus T. Pearce[3,4], and Marcos Nadal[1],

[1] Human Evolution and Cognition Research Group (EvoCog), University of the Balearic Islands, Institute for Cross-Disciplinary Physics and Complex Systems (IFISC), Associated Unit to CSIC, Palma, Spain
[2] Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain
[3] School of Electronic Engineering & Computer Science, Queen Mary University of London, UK
[4] Centre for Music in the Brain, Department of Clinical Medicine, Aarhus University, Denmark
*ana.c.magan@gmail.com

Abstract. We introduce a novel set of 200 Western tonal musical stimuli (MUST) for research on music perception and valuation. It consists of four subsets of 50 4-s motifs varying in balance, contour, symmetry, or complexity. They are musically appealing and experimentally controlled. The behavioral assessment aimed to ascertain whether musically untrained participants could identify variations in each attribute. Inter-rater reliability was high, the ratings mirrored the design features well, and the agreement served to create an abridged set. The computational assessment required developing a battery of specific computational measures that describe each stimulus in terms of its structural parameters. The distilled non-redundant composite measures proved excellent predictors of participants' ratings, and the complexity composite resulted better or as good as existing models of musical complexity. The MUST set and MATLAB toolbox constitute valuable resources for research in many fields and are freely available through OSF (https://osf.io/bfxz7/) and GitHub (https://github.com/compaes).

Keywords: music, MIR, aesthetics, multimodal, balance, contour, symmetry, complexity

Acknowledgments. The research leading to these results has received support from "la Caixa" Foundation (LCF/BQ/ES17/11600021; LCF/BQ/DE17/11600022), and grant PSI2016- 77327-P, awarded by the Spanish *Ministerio de Economía, Industria y Competitividad*. All authors reported no conflicts of interest.

Ensemble size classification in Colombian Andean string music recordings

Sascha Grollmisch[1,2], Estefanía Cano[2], Fernando Mora Ángel[3] and Gustavo López Gil[3] *

[1] Institute of Media Technology, TU Ilmenau, Ilmenau, Germany
[2] Semantic Music Technologies, Fraunhofer IDMT, Ilmenau, Germany
[3] Valores Musicales Regionales, Universidad de Antioquia, Medellín, Colombia
sascha.grollmisch@idmt.fraunhofer.de

Abstract. Reliable methods for automatic retrieval of semantic information from large digital music archives can play a critical role in musicological research and musical heritage preservation. With the advancement of machine learning techniques, new possibilities for information retrieval in scenarios where ground-truth data is scarce are now available. This work investigates the problem of counting the number of instruments in music recordings as a classification task. For this purpose, a new data set of Colombian Andean string music was compiled and annotated by expert musicologists. Different neural network architectures, as well as pre-processing steps and data augmentation techniques were systematically evaluated and optimized. The best deep neural network architecture achieved 80.7% file-wise accuracy using only feed forward layers with linear magnitude spectrograms as input representation. This model will serve as a baseline for future research on ensemble size classification.

Keywords: Ensemble Size Classification, Music Archives, Music Ensembles, Andean String Music.

1 Introduction

This work is motivated by the need of robust information retrieval techniques capable of efficiently extracting semantic information from large digital musical archives. With the advancements of deep learning techniques, numerous music information retrieval (MIR) methods have been proposed to address different information retrieval tasks, predominantly from a supervised machine learning perspective. In this work, we focus on the task of determining the size of musical ensembles, and aim to automatically classify music recordings according to the number of instruments playing in the track: solo, duet, trio, quartet, etc. Our long-term goal is to develop methods that minimally rely on manually annotated

*This work has been partially supported by the German Research Foundation (BR 1333/20-1).

data, and that can exploit commonalities between unlabeled data and the few annotations available (semi-supervised and few-shot learning). This will enable the usage of MIR techniques not only with archives of mainstream music, but also with non-western, under-represented, folk and traditional music archives. As described in section 2, not much work has been conducted on the topic of ensemble size classification in music. Consequently, this work focuses on systematically optimizing a baseline classification model in a fully supervised manner (see section 3) that can serve as a building block for future research on this topic. Detailed descriptions of the data set used and the optimization steps taken are presented in sections 3.1 and 3.2, respectively. Conclusions are presented in section 4, outlining possibilities to extend this work to semi-supervised and few-shot learning paradigms.

1.1 The ACMus Project

This research work was conducted in the context of the ACMus research project: *Advancing Computational Musicology - Semi-supervised and unsupervised segmentation and annotation of musical collections*[1]. The main goal of the project is to improve upon the limits of state-of-the-art machine learning techniques for semantic retrieval of musical metadata. In particular, ACMus focuses on leveraging semi-supervised and unsupervised techniques for segmentation and annotation of musical collections. The music collection in the *Músicas Regionales* archive at the Universidad de Antioquia in Medellín, Colombia is the focus of this research. The archive contains one of the most important collections of traditional and popular Colombian music, including music from the Colombian Andes, indigenous traditions, Afro-Colombian music, among others. The great diversity of the archive in terms of musical traditions, audio quality and formats (analogue, digital, field recordings), and musical sources (instrumental, vocal, speech, mixed), makes it a particularly challenging collection to work with. Besides developing methods for ensemble size classification, the ACMus project will also focus on developing methods for speech/music discrimination, meter recognition, and musical scale detection. The ACMus Project is a collaboration between Fraunhofer IDMT and Ilmenau University of Technology in Germany, and Universidad de Antioquia and Universidad Pontificia Bolivariana in Colombia.

2 Related work

To the best of the authors' knowledge, automatically determining the size of musical ensembles is a vastly unexplored topic in MIR research, and no state-of-the-art methods for the task have been proposed. Therefore, this section highlights source counting methods proposed in related fields such as polyphony estimation and speaker counting.

[1] `https://acmus-mir.github.io/`

2.1 Speaker Counting

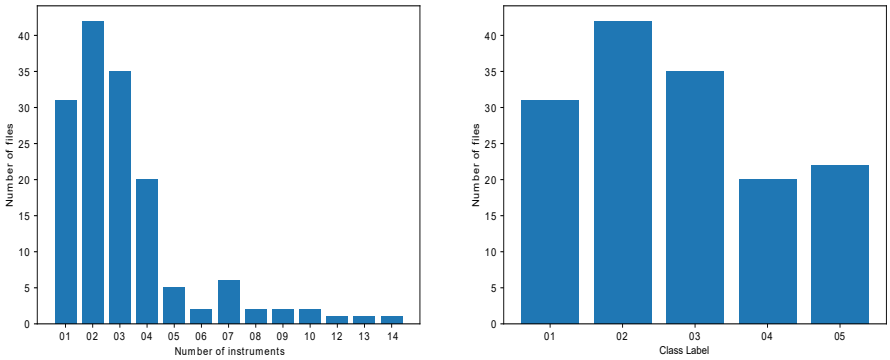
While a considerable amount of work on the topic of speaker counting for single channel recordings has been conducted, the problem has often been approached from a feature design perspective where features are specifically engineered to work with speech signals [10]. Works using more generic features such as [14][1] often assume that for the most part, only one speaker is active in the recording at a given time instant. In the case of music signals, this would be a strong assumption since musical instruments are expected to play simultaneously.
The task of audio source counting can be seen either as a regression or a classification problem when the number of maximum sources to be expected is known. In [12], the authors investigate the performance of both approaches for speaker counting using bi-directional long-short term memory neural networks (BLSTMs) with different input representations such as the linear magnitude spectrogram, the mel-scaled spectrogram, and the Mel Frequency Cepstral Coefficients (MFCCs) with linear magnitude spectrogram performing best. The data set comprised 55 hours of synthetically generated training material including signals with up to ten speakers. The system was tested on 5720 unique and unseen speaker mixtures. Even though regression could appear to be a good choice since the direct relationship of neighbouring classes is learned as well (a signal with 2 sources is closer to a signal with 3 sources than to a signal with 5), classification performed better. Based on these results, the classification approach was used in this work.

2.2 Polyphony Estimation

Polyphony estimation refers to the task of counting the number of simultaneous notes played by one or several instruments in a music recording. This can be used as a pre-processing step for multi-pitch estimation. It is important to note that polyphony estimation does not directly translate into ensemble size estimation, as several notes can be simultaneously played by a single instrument such as the guitar. Nevertheless, some relevant work on this topic is described here. Using a CNN with constant-Q transform of the audio data, the method in [2] achieved state-of-the-art performance for multi-pitch estimation. Large losses in accuracy were caused in particular by instruments playing closely harmonically related content. The authors in [6] examine this task separately with different classical instruments playing up to four simultaneous notes. Using training data of 22 minutes the proposed CNN architecture with mel-scaled spectrogram achieved a mean accuracy of 72.7% for a small evaluation set of only three songs.

3 Proposed Method for Ensemble Size Classification

Since no method has been proposed in the literature that could directly be applied to identify the number of instruments in Andean string music recordings, we focus on developing a baseline model systematically evaluated and optimized



(a) Class distribution in data set (b) Large ensembles mapped to class *5*

Fig. 1: Distribution of the annotated classes in the data set. (a) Number of files per ensemble size. (b) Final class distribution with all large ensembles mapped to class 5.

using different neural network architectures. Since neural networks achieve state-of-the-art performance in related fields, as well as in other MIR tasks such as instrument recognition [5][4], other types of supervised classifiers were not evaluated. In this study, no pre-trained models were used as we wish to build a baseline that shows the potential of different neural networks for unseen tasks, avoiding possible biases from other data sets previously used for training.

3.1 Data set

For this study, 150 representative song fragments from the *Músicas Regionales* archive were selected and annotated by at least two experts per song in Universidad de Antioquia. All the songs are instrumental pieces without vocals, performed by ensembles of plucked string instruments from the Andes region in Colombia. The instruments in the data set include different kinds of acoustic guitars, bandolas, tiples, electric bass guitars, and occasionally percussion instruments such as the maracas. The ensembles sizes considered are soloist, duet, trio, quartet, and large ensembles (five or more instruments). The annotations in the data set include the ensemble size, as well as the list of all the instruments in the ensemble.

In most songs, all annotated instruments are active during the entire file; however, short sections where one instrument is temporarily inactive also occur, leading to some instances of weak labels. The data set comprises 54 minutes of audio, with song fragment duration ranging from 7 to 62 seconds. The distribution of the classes is shown in figure 1. Songs containing five or more instruments were mapped to the class 5. No genre, composer or tempo bias was found in the class distribution. Given that the original source of the recordings include digitized versions of tape recordings as well as more recent digital recordings, these

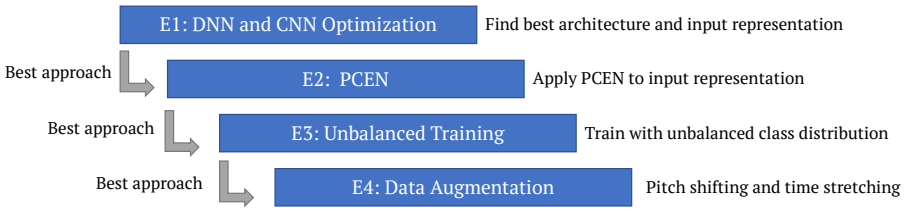


Fig. 2: Overview of the experimental setup. Four consecutive experiments (E1-E4) were performed to find the optimal architecture for our task.

files have been saved with a 96 kHz sampling rate, 24 bit-depth, and in stereo format. However, for monophonic analogue recordings, the stereo was obtained by duplicating the monophonic recording in both channels. Additionally, some of the older recordings only contain information below 8 kHz. To avoid biases during training, all files were downsampled to 12 kHz (to avoid sub-sampling artifacts), mixed to mono, and normalized to a maximum absolute amplitude of 1 for all the experiments.

3.2 Experimental setup

Four experiments were conducted in order to build a reliable baseline system, showing the upper boundaries for a fully supervised classification system with a neural network trained from scratch. As shown in figure 2, our work flow starts with Experiment 1 (E1), where different architectures and input representations are evaluated. The approach that shows best performance in E1 is then used in Experiment 2 (E2) to test the effects of per-channel energy normalization (PCEN) on the system. Similarly, E3 and E4 evaluate the effects of unbalanced training and data augmentation, respectively, on the best model from the previous experiment. In all our experiments, we performed 20 repetitions of random data set splits for testing all files and accounting for randomness during training of the networks. In each step, 70% of the files were randomly picked for training, 10% for early stopping during training, and 20% for evaluating the performance on unseen data. The test set was always balanced using the class with the smallest number of files and randomly subsampling the other classes.

Each network was trained for 500 epochs unless the validation loss stopped decreasing for 100 epochs. The Adam optimizer [7] with a learning rate of 0.001, Glorot initialization [3], categorical cross-entropy loss, and ReLU activation function (except softmax activation for the output layer), were used for all networks. For all experiments, the input representations were normalized to zero mean and standard deviation of one. The normalization values were calculated on the

training set and applied to the validation and test sets. All experiments were conducted using Tensorflow.[2]

Experiment 1 (E1) - DNN and CNN models: E1 aimed at finding the best model architecture and input representation for a feed-forward neural network (DNN), and a convolutional neural network (CNN). Bayesian Optimization [11] was used to obtain an optimal combination of hyper-parameters and comparable results for all network architectures in a reasonable amount of time.[3]

As input features, a linear magnitude spectrogram obtained from the short-time Fourier transform (STFT) was compared to the mel-scaled spectrogram with a logarithmic frequency axis (Mel) using 128 mel bands.[4] For the DNN model, the spectral frames were smoothed using a moving average filter over time for each frequency bin to highlight stable structures over several time frames while keeping the same frequency resolution and input dimensionality. The length of the averging filter, STFT size, number of layers, number of units per layer, and dropout percentage between the layers were also subject to the Bayesian optimization. For the CNN model, several time frames were combined into patches, where the patch length was also optimized. The maximum patch duration was set to 3 seconds. The basic CNN architecture was inspired by the model proposed in [5] and the number of layers and filters, amount of Gaussian noise added to input, and dropout percentage between the layers were included in the optimization. The Bayesian optimization process was performed with 30 iterations and was only feasible because of the relatively small data set (see Section 3.1).

Experiment 2 (E2) - Per-Channel Energy Normalization (PCEN): In E2, the best architectures obtained in E1 were taken, and per-channel energy normalization (PCEN)[4] was applied to each audio file. PCEN suppresses stable background noise using adaptive gain control and dynamic range compression. This has proved to be beneficial for tasks with high loudness variations such as key word spotting [13]. In this study, PCEN was applied to test its potential to account for the great variability in audio quality in our data set. PCEN was evaluated with the default settings S1 ($power = 0.5, time\_constant = 0.4, max\_size = 1$), and with a second parameter setting S2 ($power = 0.25, time\_constant = 0.01, max\_size = 20$) experimentally chosen for highlighting harmonic structures. Figure 1a and 1b show the different input representations and PCEN settings for two audio files, one with three instruments and one with four. While S1 highlights temporal changes, S2 emphasizes harmonic structures.

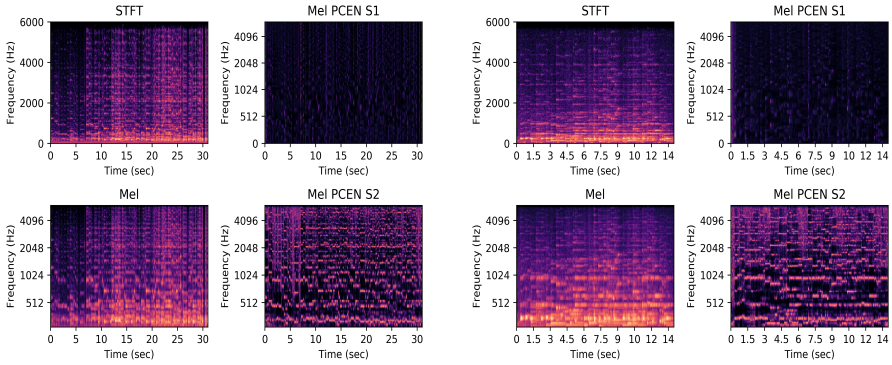
Experiment 3 (E3) - Unbalanced Training: In E1 and E2, the training data was balanced using random sub-sampling. For E3, class weights[5] were used

[2]Tensorflow (1.10): `www.tensorflow.org`

[3]Implementation from `https://github.com/fmfn/BayesianOptimization`

[4]Implementation from librosa (0.6.3): `https://librosa.github.io/`

[5]Implementation from sklearn (0.20.2): `https://scikit-learn.org/`



(a) Example with three instruments (b) Example with four instruments

Fig. 3: Input representations for two example recordings. (a) Input representations of an example of a trio. (b) Input representations of an example of a quartet.

for training on the unbalanced data set using the best CNN and DNN models from E2. Additionally, the architectures were also evaluated using the unbalanced training set without class weights in order to determine its influence on classification performance.

Experiment 4 (E4) - Data augmentation (DA): Pitch shifting and time stretching have been previously used for audio data augmentation in tasks such as chord detection [8] and singing voice separation [9]. In E4, pitch shifting (± 2 semitones), and time stretching (four steps between 90% and 110%) were applied only on the training data[4]. After data augmentation, the training set contained eight additional versions of each file.

3.3 Results

As evaluation measure, we use the mean file accuracy and standard deviation over all repetition steps. To calculate the file accuracy, the class confidences were summed up over all times frames, and the class with the highest confidence was chosen. Results are presented in Tables 1-4 and will be described in detail in the following sections. The best performing system is highlighted in bold in each table.

Experiment 1 (E1) - DNN and CNN models: Table 1 shows the results for E1. To give the reader an idea of the importance of parameter optimization, we present results for the best performing network, as well as for the worst performing one (above chance level 20%). With balanced training data and no data augmentation (E1), the highest classification accuracy (76.5%) was obtained by

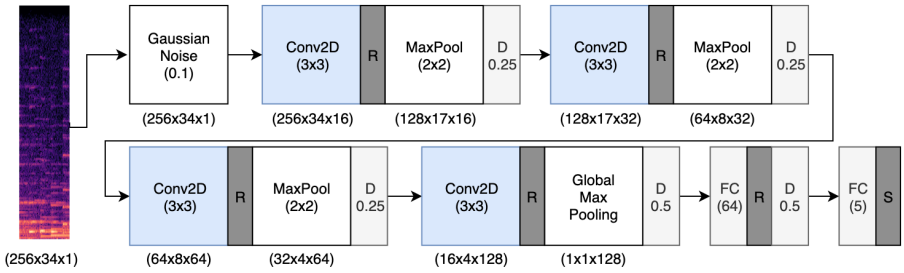


Fig. 4: Best CNN model architecture consisting of four convolutional layers (Conv2D) followed by ReLU activation (R), max pooling (MaxPool), and Dropout (D) for regularization. Global max pooling is applied before the dense layers (FC). The final dense layer uses softmax activation (S) for the classification. The corresponding output shapes are specified for each layer.

the DNN model with linear magnitude spectrogram (STFT). CNNs in general, as well as DNNs with mel-spectrogram, performed slightly worse. This suggests that with small audio training data sets, CNNs do not necessarily lead to the best performance, and that simpler and faster feed forward networks can lead to better results. Furthermore, linear magnitude spectrograms resulted in higher performance for both DNNs and CNNs. These results go in line with those reported in [12], where linear magnitude spectrogram resulted in better performance than the mel-spectrogram for speaker counting. Table 1 also shows how critical the choice of hyper-parameters is. Especially CNNs suffer when parameters are poorly chosen, leading to an accuracy of 20.5% for the worst model above chance level. Since there is so much variability in the CNNs' performance, it is possible that further optimization iterations may lead to better results and architectures than the ones found here.

Table 1: Mean accuracy, standard deviation in % for E1.

Optimization	DNN STFT	DNN Mel	CNN STFT	CNN Mel
Best	**76.5, 11.3**	72.5, 10.6	74.0, 8.7	71.3, 10.9
Worst	57.0, 14.8	65.5, 12.1	20.5, 1.6	20.5, 1.6

The final DNN model used a 2048 STFT window and hop size with logarithmic compression of the magnitudes, and a moving average filter 10 time frames long, covering in total 1.7 seconds. The 1024 unique values in the STFT were passed through a 0.1 dropout layer to one hidden layer with 512 units. The output was passed through a dropout of 0.5 to final softmax layer with 5 units, one for each class. The best CNN model is shown in detail in figure 4. The input representation was achieved from a STFT with a window and hop size of 512 samples and logarithmic compression of the magnitudes. Each patch consists of 34 STFT frames covering 1.45 seconds of audio.

Experiment 2 (E2) - PCEN: Table 2 shows the results of applying PCEN-S1 and PCEN-S2 to the input representations, as well as the best performing model from E1 (for comparison). As seen in the table, applying PCEN led to worse results when compared to E1. Between the two parameter settings of PCEN, the best results were achieved for S2 which highlights harmonic structures rather than temporal changes. In general, it appears that the suppression of possible background noise in our data when using PCEN results in the loss of discriminative information for ensemble classification. Therefore, PCEN is discarded as a processing step for the following experiments.

Table 2: Mean accuracy, standard deviation in % for E2.

PCEN	DNN STFT	DNN Mel	CNN STFT	CNN Mel
with PCEN-S1	56.0, 10.1	47.2, 9.4	60.3, 11.2	56.8, 12.3
with PCEN-S2	68.0, 10.4	67.2, 12.2	63.7, 8.3	49.2, 18.6
without PCEN (E1)	**76.5, 11.3**	72.5, 10.6	74.0, 8.7	71.3, 10.9

Experiment 3 (E3) - Unbalanced Training: Table 3 shows the results obtained with unbalanced training data, both with and without class weights. Additionally, the best performing architecture up to this point (E1) is included for comparison. The additional training data from the unbalanced training set improved the performance of all networks and lowered the standard deviation between data splits, leading to a more stable model regardless of the files chosen for training. The possible reason for the increased performance may be the increased variability of the training data since more conditions are covered in the training data set. Applying class weights led to nearly the same performance as without the weights. The reason for only having a minor impact may be that the initial data set was already nearly balanced.

Table 3: Mean accuracy, standard deviation in % for E3.

Unbalanced Training	DNN STFT	DNN Mel	CNN STFT	CNN Mel
with class weights	**80.7, 5.7**	74.8, 9.0	77.7, 7.9	73.3, 6.1
without class weights	79.7, 6.4	75.0, 9.0	77.5, 7.7	74.8, 8.8
balanced data set (E1)	76.5, 11.3	72.5, 10.6	74.0, 8.7	71.3, 10.9

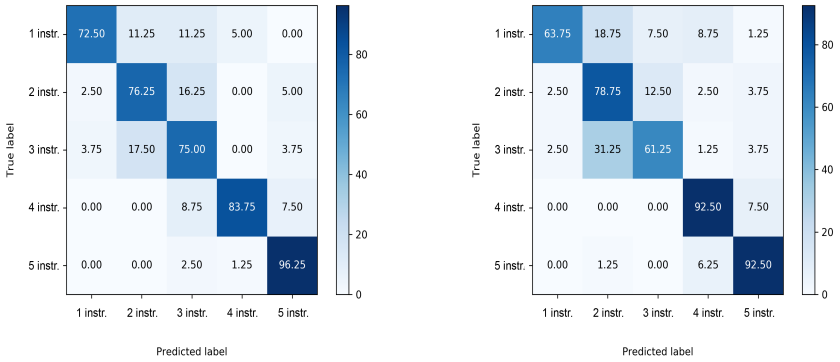
Experiment 4 (E4) - Data augmentation: Table 4 shows the results obtained with each data augmentation method, as well as the best performing architecture from E3 for comparison. Overall the best result is obtained without data augmentation using a DNN with STFT input. In contrast, the DNN model with Mel input, experiences a slight increase of performance when pitch shifting and time stretching are applied independently. Except for pure time stretching all data augmentation methods improved slightly the performance of the CNN with STFT. Using pitch shifting only led to the best CNN performance (using the STFT) with an accuracy slightly below the best DNN model. Results with the DNN go in line with those in [9] were data augmentation in small training data sets had very little impact on singing voice separation performance.

Table 4: Mean accuracy, standard deviation in % for E4.

Augmentation (DA)	DNN STFT	DNN Mel	CNN STFT	CNN Mel
with full DA	78.2, 7.7	68.5, 10.9	78.5, 6.5	77.2, 6.4
only time stretch	77.5, 7.9	78.2, 7.5	76.5, 7.1	76.2, 7.2
only pitch shift	75.2, 7.2	75.0, 9.3	80.2, 6.0	75.7, 8.2
without DA (E3)	**80.7, 5.7**	74.8, 9.0	77.7, 7.9	73.3, 6.1

3.4 Error Analysis

In order to get further insights about the classification errors of the best DNN and CNN models, figure 5 displays the mean confusion matrices for the best DNN and CNN from E3 (best overall models). Classification errors are highest between neighboring classes which shows that the network is implicitly capable of learning the relationships between classes (e.g., a duo is closer to a trio than to a quartet), and consequently, of learning useful classification features. This is in line with the findings in [6] and [12], where better performance was achieved for speaker counting with classification than with regression. It is intriguing why classification performance is relatively low for the one instrument class, which intuitively, appears to be a fairly simple classification problem. A possible explanation might be that the string instruments in our data set can simultaneously play relatively complex melodies and harmonies. This might blur the boundaries between class 1 and 2, since very similar music could alternatively be split into two different instruments. Class 5 achieved the highest classification accuracy. Since files in these class can contain up to 14 instruments, the difference between them and the other classes is probably much larger in terms of spectral content. This supports the assumption that meaningful features have been learned during training.



(a) Confusion matrix for DNN model (b) Confusion matrix for CNN model

Fig. 5: Mean confusion matrices for best models from E3.

4 Conclusions

In this work, the task of classifying the number of instruments in music recordings was addressed using a newly gathered data set of Colombian Andean string music. Apart from the challenges of the task itself, working with Andean string music comes with several difficulties: different recording conditions, scarce and expensive annotated data, and high similarity between the different instruments.

To build our baseline system, 150 tracks were annotated by expert musicologist in Colombia. Using this relatively small data set, several neural networks architectures were trained and optimized. The highest file-wise accuracy of 80.7% was achieved with a DNN, while the best CNN model attained 80.2%. Using linear magnitude spectrograms as input representation instead of its mel-scaled version, resulted in better performance in all experiments. All approaches clearly outperform the 20% chance level baseline which demonstrates the potential of this approach. In general, all networks had a minimum standard deviation of 6% between data splits, suggesting that the training set does not cover the full variance of recording conditions and instrument combinations. Neither the experiments with data augmentation using pitch shifting and time stretching nor those with PCEN showed a clear improvement in the robustness of the system. The optimization procedure showed that hyper-parameters optimization is critical when working with such a small data set. This system will serve as a baseline for future research on this topic where techniques for learning from few examples like transfer learning will be evaluated. Furthermore, techniques for incorporating unlabeled training data in a semi-supervised or unsupervised fashion will be explored.

References

1. Andrei, V., Cucu, H., Buzo, A., Burileanu, C.: Counting competing speakers in a timeframe - human versus computer. In: Interspeech Conference. ISCA, Dresden, Germany (2015)
2. Bittner, R.M., Mcfee, B., Salamon, J., Li, P., Bello, J.P.: Deep Salience Representations for F0 Estimation in Polyphonic Music. In: 18th International Society for Music Information Retrieval Conference. Suzhou, China (2017)
3. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics. Sardinia, Italy (2010)
4. Gómez, J.S., Abeßer, J., Cano, E.: Jazz Solo Instrument Classification With Convolutional Neural Networks, Source Separation, and Transfer Learning. In: 19th International Society for Music Information Retrieval Conference. Paris, France (2018)
5. Han, Y., Kim, J., Lee, K.: Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing. vol. 25, pp. 208–221 (jan 2017)
6. Kareer, S., Basu, S.: Musical Polyphony Estimation. In: Audio Engineering Society Convention 144. Milan, Italy (2018)

7. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: 3rd International Conference on Learning Representations (ICLR). San Diego, USA (2015)
8. Nadar, C.R., Abeßer, J., Grollmisch, S.: Towards CNN-based acoustic modeling of seventh chords for automatic chord recognition. In: International Conference on Sound and Music Computing. Málaga, Spain (2019)
9. Prétet, L., Hennequin, R., Royo-Letelier, J., Vaglio, A.: Singing voice separation: A study on training data. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 506–510. Brighton, UK (May 2019)
10. Sayoud, H., Boumediene, T.H., Ouamour, S., Boumediene, T.H.: Proposal of a New Confidence Parameter Estimating the Number of Speakers – An experimental investigation-. Journal of Information Hiding and Multimedia Signal Processing **1(2)**(April), 101–109 (2010)
11. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian Optimization of Machine Learning Algorithms. In: 25th International Conference on Neural Information Processing Systems. pp. 2951–2959. Lake Tahoe, Nevada (2012)
12. Stoter, F.R., Chakrabarty, S., Edler, B., Habets, E.A.P.: Classification vs. Regression in Supervised Learning for Single Channel Speaker Count Estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 436–440. IEEE (apr 2018)
13. Wang, Y., Getreuer, P., Hughes, T., Lyon, R.F., Saurous, R.A.: Trainable frontend for robust and far-field keyword spotting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5670–5674. IEEE (mar 2017)
14. Xu, C., Li, S., Liu, G., Zhang, Y.: Crowd ++ : Unsupervised Speaker Count with Smartphones Crowd ++ : Unsupervised Speaker Count with Smartphones. In: 2013 ACM international joint conference on Pervasive and ubiquitous computing. pp. 43–52. ACM, Zurich, Switzerland (2013)

Towards user-informed beat tracking of musical audio

António Sá Pinto[1,2] and Matthew E. P. Davies[1] *

[1] INESC TEC, Sound and Music Computing Group, Porto, Portugal
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
{antonio.s.pinto,matthew.davies}@inesctec.pt

Abstract. We explore the task of computational beat tracking for musical audio signals from the perspective of putting an end-user directly in the processing loop. Unlike existing "semi-automatic" approaches for beat tracking, where users may select from among several possible outputs to determine the one that best suits their aims, in our approach we examine how high-level user input could guide the manner in which the analysis is performed. More specifically, we focus on the perceptual difficulty of tapping the beat, which has previously been associated with the musical properties of expressive timing and slow tempo. Since musical examples with these properties have been shown to be poorly addressed even by state of the art approaches to beat tracking, we re-parameterise an existing deep learning based approach to enable it to more reliably track highly expressive music. In a small-scale listening experiment we highlight two principal trends: i) that users are able to consistently disambiguate musical examples which are easy to tap to and those which are not; and in turn ii) that users preferred the beat tracking output of an expressive-parameterised system to the default parameterisation for highly expressive musical excerpts.

Keywords: Beat Tracking, Expressive Timing, User Input

1 Introduction and Motivation

While the task of computational beat tracking is relatively straightforward to define – its aim being to replicate the innate human ability to synchronise with a musical stimulus by tapping a foot along with the beat – it remains a complex and unsolved task within the music information retrieval (MIR) community. Scientific progress in MIR tasks is most often demonstrated through improved accuracy scores when compared with existing state of the art methods [18]. At the core of this comparison rest two fundamental tenets: the (annotated) data upon which the algorithms are evaluated, and the evaluation method(s) used to measure performance. In the case of beat tracking, both the tasks of annotating

* António Sá Pinto and Matthew E. P. Davies are supported by Portuguese National Funds through the FCT-Foundation for Science and Technology, I.P., under the grant SFRH/BD/120383/2016 and the project IF/01566/2015.

datasets of musical material and measuring performance are non-trivial [6]. By its very nature, the concept of beat perception – how an individual perceives the beat in a piece of music – is highly subjective [15]. When tapping the beat, listeners may agree over the phase, but disagree over the tempo or preferred metrical level – with one tapping, *e.g.*, twice as fast as another, or alternatively, they may agree over the tempo, but tap in anti-phase. This inherent ambiguity led to the prevalence of multiple hypotheses of the beat, which can arise at the point of annotation, but more commonly appear during evaluation where different interpretations of ground truth annotations are obtained via interpolation or sub-sampling. In this way, a wide net can be cast in order not to punish beat tracking algorithms which fail to precisely match the annotated metrical level or phase of the beats; with this coming at the expense that some unlikely beat outputs may inadvertently be deemed accurate. Following this evaluation strategy, the performance of the state of the art is now in the order of 90% on existing datasets [3, 4] comprised primarily of pop, rock and electronic dance music. However, performance on more challenging material [10] is considerably lower, with factors such as expressive timing (*i.e.*, the timing variability that characterises a human performance, in opposition to a metronomic or perfectly timed rendition [7]), recording quality, slow tempo and metre changes among several identified challenging properties.

Although beat tracking has garnered much attention in the MIR community, it is often treated as an element in a more complex processing pipeline which provides access to "musical time", or simply evaluated based on how well it can predict ground truth annotations. Yet, within the emerging domain of creative-MIR [16, 11] the extraction of the beat can play a critical role in musically-responsive and interactive systems [13]. A fundamental difference of applying beat tracking in a creative application scenario is that there is a specific end-user who wishes to directly employ the music analysis and thus has very high expectations in terms of its performance [1]. To this end, obtaining high mean accuracy scores across some existing databases is of lower value than knowing "*Can the beats be accurately extracted (as I want them) for this specific piece of music?*". Furthermore, we must also be aware that accuracy scores themselves may not be informative about "true" underlying performance [17, 6].

Of course, a user-specific beat annotation can be obtained without any beat tracking algorithm, by manually annotating the desired beat locations. However, manually annotating beat locations is a laborious procedure even for skilled annotators [10]. An alternative is to leverage multiple beat interpretations from a beat tracking algorithm, and then provide users with a range of solutions to choose from [8]. However, even with a large number of interpretations (which may be non-trivial and time-consuming to rank) there is no guarantee that the end-user's desired result will be present, especially if the alternative interpretations are generated in a deterministic manner from a single beat tracking output, *e.g.*, by interpolation or sub-sampling.

In this paper, we propose an alternative formulation which allows an end-user to drive how the beat tracking is undertaken. Our goal is to enable the

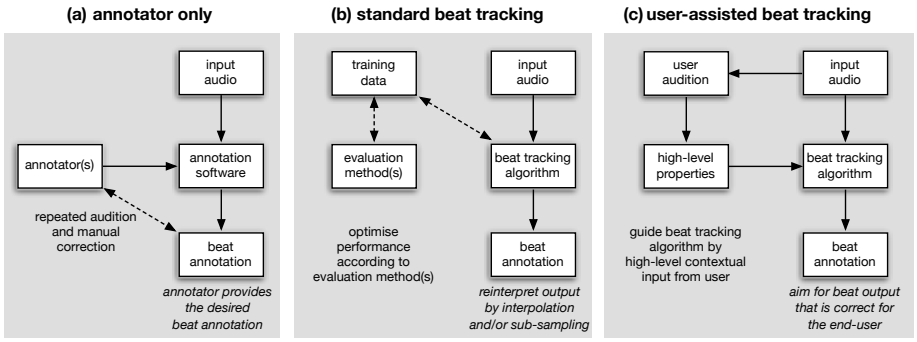


Fig. 1. Overview of different approaches to obtaining a desired beat annotation. (a) The user annotates the beat positions. (b) A beat tracking algorithm is used – whose performance has been optimised on annotated datasets. (c) Our proposed approach, where user input guides the beat tracking.

user to rapidly arrive at the beat annotation suitable for their purposes with a minimal amount of interaction. Put another way, we envisage an approach to beat tracking where high-level contextual knowledge about a specific musical signal can be given by the user and reliably interpreted by the algorithm, without the need for extensive model training on annotated datasets, as shown in Fig. 1. In this sense, we put aside the concept of "universal" beat tracking models which target equal performance irrespective of the musical input signal, in favour of the more realistic goal of identifying different classes of the beat tracking problem, which require different beat tracking strategies. While the end goal of retrieving beat locations may be the same for fast-paced techno music and highly expressive classical guitar recordings, the assumptions about what constitutes the beat, and how this can be extracted from audio signals are not. Conversely, constraints should not be placed on what musical content can be creatively re-purposed based on the limitations of MIR algorithms.

The long term challenges of our approach are as follows: i) determining a low-dimensional parameterisation of the beat tracking space within which diverse, accurate solutions can be found in order to match different beat tracking conditions; ii) exposing these dimensions to end-users in a way that they can be easily understood; iii) providing an interpretable and understandable mapping between the user-input and the resulting beat annotation via the beat tracking algorithm; and finally iv) measuring the level of engagement among end-users who actively participate in the analysis of music signals.

Concerning the dimensions of beat tracking, it is well-understood that music of approximately constant (medium) tempo, with strong percussive content (*e.g.*, pop, rock music) is straightforward to track. Beat tracking difficulty (both for computational approaches and human tappers) can be due to musical reasons and signal-based properties [9, 10]. While it is somewhat nonsensical to consider a piece of music with "opposite" properties to the most straightforward case, it has

been shown empirically that highly expressive music, without clear percussive content, is not well analysed even by the state of the art in beat tracking [10, 4]. Successful tracking of such pieces should, in principle, require input features which can be effective in the absence of percussion and a tracking model which can rapidly adapt to expressive tempo variation. While recent work [3] sought to develop multiple beat tracking models, these were separately trained at the level of different databases rather than according to musical beat tracking conditions.

In our approach, we reexamine the functionality of the current state of the art in beat tracking, *i.e.*, the recurrent neural network approach of Böck et al. [4]. In particular, we devise a means to re-parameterise it so that it is adapted for highly expressive music. Based on an analysis of existing annotated datasets, we identify a set of musical stimuli we consider typical of highly challenging conditions, together with a parallel set of "easier" examples. We then conduct a small-scale listening experiment where participants are first asked to rate the perceptual difficulty of tapping the beat, and subsequently to rate the subjective quality of beat annotations given by the expressive parameterisation vs the default version. Our results indicate that listeners are able to distinguish easier from more challenging cases, and furthermore that they preferred the beat tracking output of the expressive-parameterised system to the default parameterisation for the highly expressive musical excerpts. In this sense, we seek to use the assessment of perceptual difficulty of tapping as a means to drive the manner in which the beats can be extracted from audio signals towards the concept of user-informed beat tracking. To complement our analysis, we explore the objective evaluation of the beat tracking model with both parameterisations.

The remainder of this paper is structured as follows. In Section 2 we detail the adaption of the beat tracking followed by the design of a small-scale listening experiment in Section 3. This is followed by results and discussion in Section 4, and conclusions in Section 5.

2 Beat Tracking System Adaptation

Within this work our goal is to include user input to drive how music signal analysis is conducted. We hypothesise that high-level contextual information which may be straightforward for human listeners to determine can provide a means to guide how the music signal analysis is conducted. For beat tracking, we established in Section 1 that for straightforward musical cases, the current state of the art [4] is highly effective. Therefore, in order to provide an improvement over the state of the art, we must consider the conditions in which it is less effective, in particular those displaying expressive timing. To this end, we first summarise the main functionality of the beat tracking approach of Böck et al., after which we detail how we adapt it.

The approach of Böck et al. [4] uses deep learning and is freely available within the madmom library [2]. The core of the beat tracking model is a recurrent neural network (RNN) which has been trained on a wide range of annotated beat tracking datasets to predict a beat activation function which exhibits peaks at

likely beat locations. To obtain an output beat sequence, the beat activation function given by the RNN is post-processed by a dynamic Bayesian network (DBN) which is approximated by a hidden Markov model [14].

While it would be possible to retain this model from scratch on challenging data, this has been partially addressed in the earlier multi-model approach of Böck et al. [3]. Instead, we reflect on the latter part of the beat tracking pipeline, namely how to obtain the beat annotation from the beat activation function. To this end, we address three DBN parameters: i) the minimum tempo in beats per minute (BPM); ii) the maximum tempo; and iii) the so-called "transition-λ" parameter which controls the flexibility of the DBN to deviate from a constant tempo[3]. Through iterative experimentation, including both objective evaluation on existing datasets and subjective assessment of the quality of the beat tracking output, we devised a new set of expressiveness-oriented parameters, which are shown, along with the default values in Table 1. More specifically, we first undertake a grid search across these three parameters on a subset of musical examples from existing annotated datasets for which the state of the art RNN is deemed to perform poorly, *i.e.*, by having an information gain lower than 1.5 bits [19]. An informal subjective assessment was then used to confirm that reliable beat annotations could be obtained from the expressive parameterisation.

Table 1. Overview of default and expressive-adapted parameters.

Parameter	**Default**	**Expressive**
Minimum Tempo (BPM)	55	35
Maximum Tempo (BPM)	215	135
Transition-λ (unitless)	100	10

As shown in Table 1, the main changes for the expressive model are a shift towards a slower range of allowed tempi (following evidence about the greater difficulty of tapping to slower pieces of music [5]), together with a lower value for the transition-λ. While the global effect of the transition-λ was studied by Krebs et al. [14], their goal was to find an optimal value across a wide range of musical examples. Here, our focus is on highly expressive music and therefore we do not need a more general solution. Indeed, the role of the expressive model is to function in precisely the cases where the default approach can not.

3 Experimental Design

Within this paper, we posit that high-level user-input can lead to improved beat annotation over using existing state of the art beat tracking algorithms in a

[3] the probability of tempo changes varies exponentially with the negative of the "transition-λ", thus higher values of this parameter favour constant tempo from one beat to the next one [14].

"blind" manner. In order to test this in a rigorous way, we would need to build an interactive beat tracking system including a user interface, and conduct a user study in which users could select their own input material for evaluation. However, doing so would require understanding which high-level properties to expose and how to meaningfully interpret them within the beat tracking system. To the best of our knowledge, no such experiment has yet been conducted, thus in order to gain some initial insight into this problem, we conducted a small-scale online listening experiment, which is split into two parts: **Part A** to assess the perceptual difficulty of tapping the beat, and **Part B** to assess the subjective quality of beat annotations made using the default parameterisation of the state of the art beat tracking system versus our proposed expressive parameterisation.

We use **Part A** as a means to simulate one potential aspect of high-level context which an end-user could provide: in this case, a choice over whether the piece of music is easy or difficult to tap along to (where difficulty is largely driven by the presence of expressive timing). Given this choice, **Part B** is used as the means for the end-user to rate the quality of the beat annotation when the beat tracking system has been parameterised according to their choice. In this sense, if a user rates the piece as "easy", we would provide the default output of the system, and if they rate it as "hard" we provide the annotation from the expressive parameterisation. However, for the purposes of our listening experiment, all experimental conditions are rated by all participants, thus the link between **Part A** and **Part B** is not explicit.

3.1 Part A

In the first part of our experiment, we used a set of 8 short music excerpts (each 15 s in duration) which were split equally among two categories: i) "easy" cases with near constant tempo in 4/4 time, with percussive content, and without highly syncopated rhythmic patterns; and ii) "hard" cases typified by the presence of high tempo variation and minimal use of percussion. The musical excerpts were drawn from existing public and private beat tracking datasets, and all were normalised to -3 dB.

We asked the participants to listen to the musical excerpts and to spontaneously tap along using the computer keyboard at what they considered the most salient beat. Due to the challenges of recording precise time stamps without dedicated signal acquisition hardware (*e.g.*, at the very least, a MIDI input device) the tap times of the participants were not recorded, however this was not disclosed. We then asked the participants to rate the difficulty they felt when trying to tap the beat, according to the following four options:

- Low - *I could easily tap the beat, almost without concentrating*
- Medium - *It wasn't easy, but with some concentration, I could adequately tap the beat*
- High - *I had to concentrate very hard to try to tap the beat*
- Extremely high - *I was not able to tap the beat at all.*

Our hypothesis for **Part A** is that participants should consistently rate those drawn from the "easy" set as having Low or Medium difficulty, whereas those from the "hard" should be rated with High or Extremely High difficulty.

3.2 Part B

Having completed **Part A**, participants then proceeded to **Part B** in which they were asked to judge the subjective quality of beat annotations (rendered as short 1 kHz pulses) mixed with the musical excerpts. The same set of musical excerpts from **Part A** were used, but they were annotated in three different ways: i) using the *default* parameterisation of the Böck et al. RNN approach from the madmom library [2]; ii) using our proposed *expressive* parameterisation (as in Table 1); and iii) a control condition using a completely *deterministic* beat annotation, *i.e.*, beat times at precise 500 ms intervals without any attempt to track the beat of the music. In total, this created a set of $8 \times 3 = 24$ musical excerpts to be rated, for which participants were asked to: *Rate the overall quality of how well the beat sequence corresponds to the beat of the music.*

For this question, a 5-point Likert-type item was used with (1) on the left hand side corresponding to "Not at all" and (5) corresponding to "Entirely" on the right hand side. Our hypothesis for **Part B** was that for the "hard" excerpts, the annotations of the expressively-parameterised beat tracker would be preferred to those of the default approach, and for all musical excerpts that the deterministic condition would be rated the lowest in terms of subjective quality.

3.3 Implementation

The experiment was built using HTML5 and Node.js and run online within a web browser, where participants were recruited from the student body of the University of Porto and the research network of the Sound and Music Computing Group at INESC TEC. Within the experimental instructions, all participants were required to give their informed consent to participate, with the understanding that any data collected would be handled in an anonymous fashion and that they were free to withdraw at any time without penalty (and without their partial responses being recorded). Participants were asked to provide basic information for statistical purposes: sex, age, their level of expertise as a musician, and experience in music production.

All participants were encouraged to take the experiment in a quiet environment using high quality headphones or loudspeakers, and before starting, they were given the opportunity to set the playback volume to a comfortable level. Prior to the start of each main part of the experiment, the participants undertook a compulsory training phase in order to familiarise themselves with the questions. To prevent order effects, each participant was presented with the musical excerpts in a different random order. In total, the test took around 30 minutes to complete.

4 Results and Discussion

4.1 Listening Experiment

A total of 10 listeners (mean age: 31, age range: 23–43) participated in the listening test, 9 of whom self-reported amateur or professional musical proficiency.

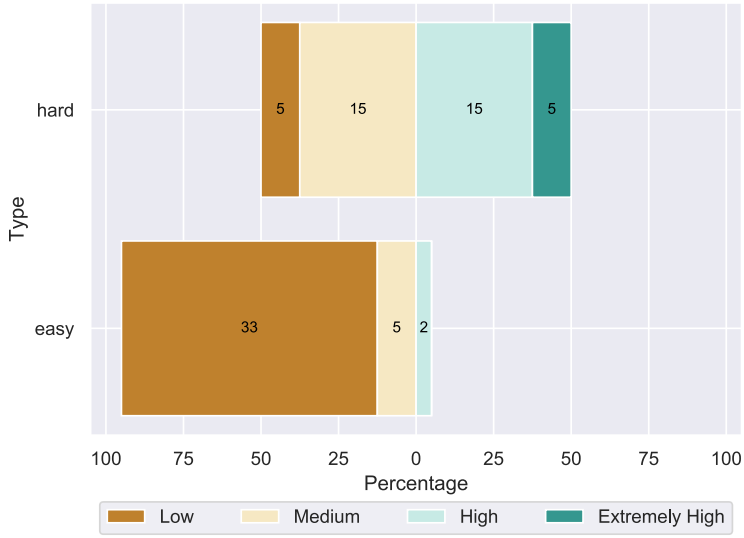


Fig. 2. Subjective ratings of the difficulty of beat tapping.

For **Part A**, we obtained 40 ratings for each stimuli group "easy" and "hard", according to the frequency distribution shown in Fig. 2. The most frequent rating for the first group was "low" (82.5%), followed by the "medium" rating (12.5%). For the "hard" group, a symmetrical rating was obtained: the adjacent ratings "medium" and "high" (37.5% each), complemented by the more extreme ratings "low" and "extremely high" (12.5% each). A Mann-Whitney test showed that there was a statistically significant difference between the ratings for both groups, with $p < 0.001$.

From these results we interpret that there was greater consistency in classifying the "easy" excerpts as having low difficulty, with only two excerpts rated above "medium", than for the "hard" excerpts which covered the entire rating scale from low to extremely difficult, albeit with the majority of ratings being for medium or high difficulty. We interpret this greater variability in the rating of difficulty of tapping to be the product of two properties of the participants: their expertise in musical performance and/or their familiarity with the pieces. Moreover, we can observe a minor separation between the understanding of the

perceptual difficulty in tapping on the part of the participant and the presence of expressive timing in the musical excerpts; that experienced listeners may not have difficulty in tapping along with a piece of expressive music for which they knew well. Thus, for expert listeners it may be more reasonable to ask a direct question related to the presence of expressive timing, while the question of difficulty may be more appropriate for non-expert listeners who might lack familiarity with the necessary musical terminology.

For **Part B**, we again make the distinction between the ratings of the "easy" and the "hard" excerpts. A Kruskal-Wallis H test showed that there was a statistically significant difference between the 3 models (*expressive*, *default* and *deterministic*): $\chi^2(2) = 87.96$, $p < 0.001$ for "easy" excerpts, $\chi^2(2) = 70.71$, $p < 0.001$ for "hard" excerpts. A post-hoc analysis performed with the Dunn test with Bonferroni correction showed that all the differences were statistically significant with $p < 0.001/3$ (except for the pair *default–expressive* under the "easy" stimuli, for which identical ratings were obtained). A descriptive summary of the ratings (boxplot with scores overlaid) for each type of stimuli, and under the three beat annotation conditions are shown in Fig. 3.

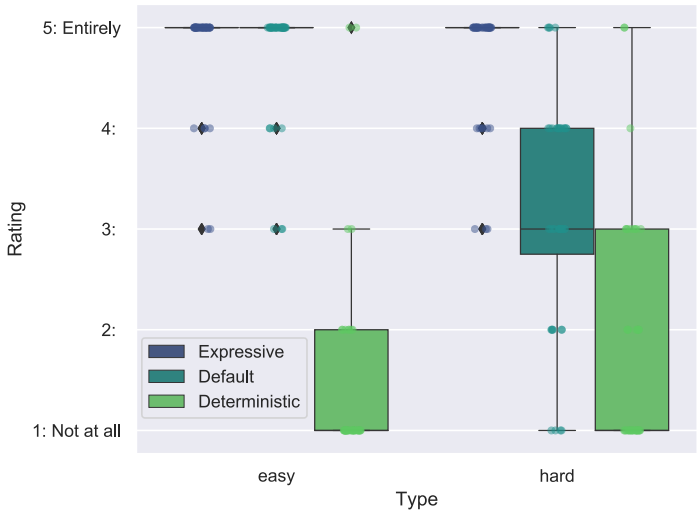


Fig. 3. Subjective ratings of the quality of the beat annotations.

The main results from Part B are as **follows. For** the "easy" excerpts there is no difference in performance for the *default* and *expressive* parameterisations of the beat tracking model, both of which are rated with high scores indicating high quality beat annotations from both systems. We contrast this with the ratings of the *deterministic* output (which should bear no meaningful relationship to the music) and which are rated toward the lower end of the scale. From these

results we can infer that the participants were easily able to distinguish accurate beat annotations and entirely inaccurate annotations, which is consistent with the Beat Alignment Test [12]. Concerning the ability of the expressively parameterised model to achieve such high ratings, we believe that this was due to very clear information concerning the beat in the beat activation functions from the RNN.

Conversely, the ratings of the "hard" excerpts show a different picture. Here, the ratings of the expressively parameterised model are similar to the "easy" excerpts, but the ratings of the *default* model [2] are noticeably lower. This suggests that the participants, in spite of their reported higher perceptual difficulty in tapping the beat, were able to reliably identify the accurate beat predictions of the *expressive* model over those of the *default* model. It is noteworthy that the ratings of the *deterministic* approach are moderately higher for the "hard" excerpts compared to the "easy" excerpts. Given the small number of samples and participants for this experiment, it is hard to draw strong conclusions about this difference, but for highly expressive pieces, the *deterministic* beats may have inadvertently aligned with the music in brief periods compared to the "easy" excerpts, which may have been unambiguously unrelated.

4.2 Beat Tracking Accuracy

In addition to reporting on the listening experiment whose focus is on subjective ratings of beat tracking, we also examine the difference in objective performance of using the *default* and *expressive* parameterisations of the beat tracking model. Given the focus on challenging excerpts for beat tracking, we focus on the SMC dataset [10]. It contains 217 excerpts, each of 40 s in duration. Following the evaluation methods described in [6] we select a common subset: F-measure, CMLc, CMLt, AMLc, AMLt, and the Information Gain (D) to assess performance. In Table 2, we show the recorded accuracy on this dataset for both the default and expressive parameterisations. Note, for the default model we use the version in the madmom library [2] which has been exposed to this material during training, hence the accuracy scores are slightly higher than those in [4] where cross fold validation was used. In addition to showing the performance of each parameterisation we also show the theoretical upper limit achievable by making a perfect choice (by a hypothetical end-user) among the two parameterisations.

Table 2. Overview of beat tracking performance on the SMC dataset [10] comparing the default and expressive parameters together with upper limit on performance.

	F-measure	CMLc	CMLt	AMLc	AMLt	D
Default[2]	0.563	0.350	0.472	0.459	0.629	1.586
Expressive	0.540	0.306	0.410	0.427	0.565	1.653
Optimal Choice	0.624	0.456	0.611	0.545	0.703	1.830

From Table 2, we see that for all the evaluation methods, with the exception of the Information Gain (D), the default parameterisation outperforms the expressive one. This is an expected result since the dataset is not entirely comprised of highly expressive musical material. We consider the more important result to be the potential for our *expressive* parameterisation to track those excerpts for which the *default* approach fails. To this end, the increase of approximately 10% points across each of the evaluation methods demonstrates how these two different parameterisations can provide greater coverage of the dataset. It also implies that training a binary classifier to choose between expressive and non-expressive pieces would be a promising area for future work.

5 Conclusions

In this paper we have sought to open the discussion about the potential for user-input to drive how MIR analysis is performed. Within the context of beat tracking, we have demonstrated that it is possible to reparameterise an existing state-of-the-art approach to provide better beat annotations for highly expressive music, and furthermore, that the ability to choose between the default and expressive parameterisation can provide significant improvements on very challenging beat tracking material. We emphasise that the benefit of the expressive model was achieved without the need for any retraining of the RNN architecture, but that the improvement was obtained by reparameterisation of the DBN tracking model.

To obtain some insight into how user input could be used for beat tracking, we simulated a scenario where user decisions about perceptual difficulty of tapping could be translated into the use of a parameterisation for expressive musical excerpts. We speculate that listener expertise as well as familiarity may play a role in lowering the perceived difficulty of otherwise challenging expressive pieces. Our intention is to further investigate the parameters which can be exposed to end-users, and whether different properties may exist for expert compared to non-expert users. Despite the statistical significance of our results, we recognise the small-scale nature of the listening experiment, and we intend to expand both the number of musical excerpts uses as well as targeting a larger group of participants to gain deeper insight into the types of user groups which may emerge. Towards our long-term goal, we will undertake an user study not only to understand the role of beat tracking for creative MIR, but also to assess the level of engagement when end-users are active participants who guide the analysis.

References

1. K. Andersen and P. Knees. Conversations with Expert Users in Music Retrieval and Research Challenges for Creative MIR. In *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, pages 122–128, 2016.
2. S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. Madmom: A new python audio and music signal processing library. In *Proc. of the 2016 ACM Multimedia Conf.*, pages 1174–1178, 2016.

3. S. Böck, F. Krebs, and G. Widmer. A multi-model approach to beat tracking considering heterogeneous music styles. In *Proc. of the 15th Intl. Society for Music Information Retrieval Conf.*, pages 603–608, 2014.
4. S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, pages 255–261, 2016.
5. R. Bååth and G. Madison. The subjective difficulty of tapping to a slow beat. In *Proc. of the 12th Intl. Conf. on Music Perception and Cognition*, pages 82–55, 2012.
6. M. E. P. Davies and S. Böck. Evaluating the evaluation measures for beat tracking. In *Proc. of the 15th Intl. Society for Music Information Retrieval Conf.*, pages 637–642, 2014.
7. P. Desain and H. Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56(4):285–292, 1994.
8. M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. Songle: A Web Service for Active Music Listening Improved by User Contributions. In *Proc. of the 12th Intl. Society for Music Information Retrieval Conf.*, pages 311–316, 2011.
9. P. Grosche, M. Müller, and C. Sapp. What makes beat tracking difficult? a case study on chopin mazurkas. In *Proc. of the 11th Intl. Society for Music Information Retrieval Conf.*, pages 649–654, 2010.
10. A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(9):2539–2460, 2012.
11. E. J. Humphrey, D. Turnbull, and T. Collins. A brief review of creative MIR. In *Late-breaking demo session of the 14th Intl. Society for Music Information Retrieval Conf.*, 2013.
12. J. R. Iversen and A. D. Patel. The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In *Proc. of the 10th Intl. Conf. on Music Perception and Cognition*, pages 465–468, 2010.
13. C. T. Jin, M. E. P. Davies, and P. Campisi. Embedded Systems Feel the Beat in New Orleans: Highlights from the IEEE Signal Processing Cup 2017 Student Competition [SP Competitions]. *IEEE Signal Processing Magazine*, 34(4):143–170, 2017.
14. F. Krebs, S. Böck, and G. Widmer. An efficient state space model for joint tempo and meter tracking. In *Proc. of the 16th Intl. Society for Music Information Retrieval Conf.*, pages 72–78, 2015.
15. D. Moelants and M. McKinney. Tempo perception and musical content: what makes a piece fast, slow or temporally ambiguous? In *Proc. of the 8th Intl. Conf. on Music Perception and Cognition*, pages 558–562, 2004.
16. X. Serra et al. Roadmap for music information research, 2013. Creative Commons BY-NC-ND 3.0 license, ISBN: 978-2-9540351-1-6.
17. B. L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.
18. J. Urbano, M. Schedl, and X. Serra. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369, 2013.
19. J. R. Zapata, A. Holzapfel, M. E. P. Davies, J. L. Oliveira, and F. Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proc. of the 13th Intl. Society for Music Information Retrieval Conf.*, pages 157–162, 2012.

Drum Fills Detection and Generation

Frederic Tamagnan and Yi-Hsuan Yang *

Academia Sinica, Taiwan
frederic.tamagnan@gmail.com
yang@citi.sinica.edu.tw

Abstract. Drum fills are essential in the drummer's playing. They regularly restore energy and announce the transition to a new part of the song. This aspect of the drums has not been explored much in the field of MIR because of the lack of datasets with drum fills labels. In this paper, we propose two methods to detect drum fills along a song, to obtain drum fills context information. The first method is a logistic regression which uses velocity-related handcrafted data and features from the latent space of a variational autoencoder. We give an analysis of the classifier performance regarding each features group. The second method, rule-based, considers a bar as a fill when a sufficient difference of notes is detected with respect to the adjacent bars. We use these two methods to extract regular pattern/ drum fill couples in a big dataset and examine the extraction result with plots and statistical test. In a second part, we propose a RNN model for generating drum fills, conditioned by the previous bar. Then, we propose objective metrics to evaluate the quality of our generated drum fills, and the results of a user study we conducted. Please go to https://frederictamagnan.github.io/drumfills/ for details and audio examples.

Keywords: Drum fills detection, Drum fills generation

1 Introduction

Percussions and drums are a fundamental core aspect of music. One important part of long-term drums generation is the drum fills issue. In recent works on music generation using generative deep learning models, drum fills have often been treated implicitly. The main challenge of drum fills generation comes from the lack of labelled data. So that, drum fills detection is an important preliminary task. The second tricky issue that comes in mind when dealing with drum fills is the lack of rules that defined them. Nevertheless, our empirical observations can lead to these properties: 1) a greater use of toms, snares or cymbals, than in the regular drum pattern; 2) a difference of played notes between the regular pattern and the drum fills; 3) an appearance in general at the end of a cycle of 4 or 8 bars.The task of detecting and generating drum fills explicitly has at least the following two use cases : first, segmenting the parts of a music piece,

* This work was done when FT was a visiting student at Academia Sinica.

as important drum fills are often located as a transition between two parts of a song, from the verse to the chorus for example; second, allowing the generation of long music sequences, in order to be able to create drum patterns with real evolution and ruptures.
In this paper, we present an initial attempt towards generating drum fills. Our goal is first to address drum fills detection and to build-up a dataset of regular pattern/drum fills couples (Figure 1). Secondly, we use this dataset to train a model able to generate a drum fill based on a regular pattern. In particular, this work allows us to answer three research questions: (1) Can we train a fill detector from isolated fills? or is it mandatory to take into account the context? (2) Is a rule-based method effective enough to detect fills? (3) How objectively a human can rate a drum fill? In sections 4–5, we develop two methods to detect and classify drum fills. The first is a logistic regression based on two different group of features: velocity-related handcrafted features and variables from a variational auto-encoder latent space. The classifier has been trained on drums kits from Native Instruments and OddGrooves.com with regular pattern and drum fills labels. The second method is a rule-based method that reflects the interpretation of a drum fill as a variation. Then, in Section 6 using these two classifiers, we extract regular pattern/ drum fills couples in the Lakh pianoroll dataset to build-up two generation datasets. After cleaning these extracted datasets to provide clean and balanced enough datasets to our further generation model, we evaluate the extraction. Our generation model, whose architecture is precisely described in Section 7, is able to generate a drum fill based on the regular pattern given as a input. We use a many-to-many RNN with 2 layers of GRU units, followed by fully-connected and batch-normalization layers. Section 8 shows the results of the user-study we have conducted with musicians and non musicians where our model trained on our two different datasets is confronted with a rule-based method to generate drum fills.

2 Related Works

Lopez-Serrano *et al.*. [5] have proposed a method to detect drum breaks in the audio domain. In this paper it is not a question of detecting short drum fills but rather percussion-only passages. The authors address this problem inspired by a baseline method initially designed for singing voice detection. In order to detect frames that contain percussion-only passages, they use features in the audio domain as included in [6], and a random forest model to define a median filtered decision function over the frames, and then apply a decision threshold. Roberts *et al.* [3], wrote a paper about learning and generating long term structure music. Recurrent Variationnal Auto-Encoder (VAE) having difficulties to model a piece of music made up of several bars, they use a VAE including a hierarchical decoder. The VAE encoder produces a latent vector from n bars using a bi-directional recurrent layer. The first level of the decoder, the conductor, generates a series of embedding vectors from the latent vector, each corresponding to a bar. A second level of recurrent decoder decodes these embeddings into

notes sequences. The most interesting thing in this paper related to our topic, is that their model is able to produce drum fills implicitly as we can hear in their demo files.

3 Preliminaries

In this paper, we do not care about the precise boundaries of a drum fills. To simplify the problem, we decide to detect and generate bars containing drum fills. We also reduce the problem by working with only 9 different drums instruments as [3]: kick (abbreviated BD for bass drum), snare (SD), low, mid and high tom (LT,MT and HT), closed and open hi-hat (CHH and OHH), crash and ride cymbals (CC and RC). We work only with bars having a 4/4 time signature. We decide to work with a precision of 4 time steps for each beat. This gives us a tensor with a 9×16 dimensions filled with the velocity of each note. In the next sections, we use the term "reduced pianoroll" to call a pianoroll transformed to a 9×16 tensor and "binarized pianoroll" to call a pianoroll filled with 0 and 1 instead of velocity.

3.1 Datasets

Labelled Datasets The Native Instruments' Battery Kits and the oddgrooves website's fill pack are composed by loops with different time signatures and length. We decided to crop and add paddings to these loops to form bars with a 4/4 signature. The concatenation of these bars from the two datasets gives us a dataset composed of 5,317 regular patterns and 1,412 drum fills.

Unlabelled Dataset The dataset we would like to label is the Lakh Pianoroll Dataset [2], a derivative of Lakh Midi Dataset [1], which contains 21,425 songs with their related drums pianorolls.

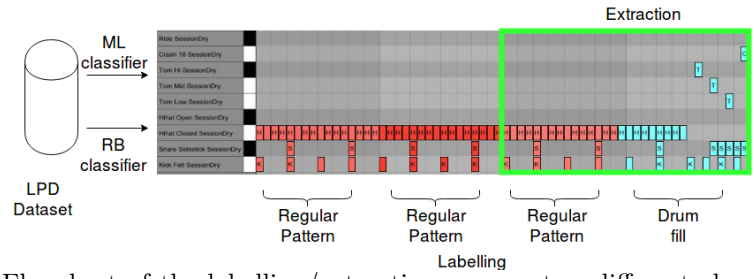


Fig. 1: Flowchart of the labelling/extraction: we use two different classifiers to label the LPD dataset. Then, we extract regular pattern/drum fill couples to constitute a drum fills dataset for generation

4 Machine Learning Classifier

4.1 Features Used and Model

We use two groups of features to train our model. For each bar of our labelled dataset, we decide to compute the maximum, the standard deviation and the mean of the velocity for each instrument of the reduced drums classes along the time axis. It give us a 27 dimensions-vector. This vector represents the use, the amount of notes and the dynamics of playing of each drum class for each bar. We trained a VAE over thousands of bars of the Lakh pianoroll dataset [2], to obtain features that capture a good compressed representation of the drums patterns. Then, we use the encoder of this VAE to encode the data of our labelled dataset and to obtain the latent space features. It gives us a 32-dimensions vector.

We train a logistic classifier with regularization on our whole labelled dataset using *LogisticRegressionCV* from the Sklearn API [7]. We use standardization as pre-processing of our data and automatic-cross validation to tune the regularization hyperparameter.

4.2 Validation

Using the $L2$ regularization that performs better in our case, we obtain the result shown in Table 1.

Feature set	Precision	Recall	F1 Score
HD	0.80	0.79	0.79
LS	0.58	0.06	0.10
HD+LS	0.89	0.81	0.85

Table 1: Validation metrics of our classifier. HD: Handcrafted features, LS: VAE's latent space features

The results for the VAE's latent space features are low because there were few drum fills compared to regular patterns in the VAE's training dataset. So that, latent space features badly capture the essence of drum fills. Although the training with $L1$ regularization has worse performance results, it is interesting to have a look on the weights, to see which features are the most correlated with the purpose of detecting fills. The three most correlated LS features are the 18th, 20th and 1st latent space variables ; they are associated with the regression coefficients 2.06,1.92 and 1.61. The three most correlated HD features are the max velocity of high tom, the standard deviation of mid tom and the max velocity of low tom ; they are associated with the regression coefficients 1.26, 1.26 and 1.26. That confirms our intuition that fills are related with toms and cymbals and that gives us a better comprehension of our VAE. The drawback of this approach is that we characterize a drum fill with absolute rules, and not with the relative difference between bars.

5 Rule-based Classifier

Fills can be seen as a variation regarding the regular pattern of the song they belong to. In order to answer to research question 2, we build another approach, rule-based. Let A, B be two binarized pianorolls (tensors) of dimension $t \times n$ (time steps $\times$ number of instruments), we define the difference of notes DN between A and B as:

$$DN(A, B) = \sum_{\substack{0 \leq i < t \\ 0 \leq j < n}} max(0, A_{i,j} - B_{i,j}) \tag{1}$$

Iterating over the binarized and reduced bars of our unlabelled dataset, we decide to consider the current bar as a drum fill if the difference of notes between the current bar and the two adjacents bars respectively is above a threshold. We use a threshold of 7 notes for the extraction part.

6 Extraction of Fills

We apply our machine learning classifier and our rule-based classifier to our unlabelled dataset. Then, we extract 2-bars sequences composed by a drum fill following a regular pattern. So, we obtain two datasets from our two labelling methods that we will call ML dataset (extracted with the machine learning classifier) and RB dataset (extracted with the Rule-based classifier).

6.1 Data Cleaning

In order to have a good enough datasets for the generation we apply the three following rules (Table 2) to clean our datasets: removing duplicated rows (Rule 1), removing all the couples where the regular pattern or the drum fill have fewer than 7 notes (Rule 2), removing all the couple where the drum fill has a too high density of snare notes, above 8 (Rule 3) .

	#ML dataset	#RB dataset
Raw	13,476	97,023
After Rule 1	6,324	45,723
After Rule 2	5,271	39,108
After Rule 3	3,283	32,130

Table 2: Influence of the cleaning process on our datasets size. The RB dataset is less sensitive to our filtering rules.

6.2 Analysis of the Extracted Datasets

Total of Notes by Instrument A Pearson's chi-squared test between the total of notes by instrument of the regular patterns and the drum fills certifies us that the distributions are significantly different for the two extracted datasets. The drum fills of the two datasets contains more toms and cymbals notes than the regular patterns, as we can see for example in the Figure 2.

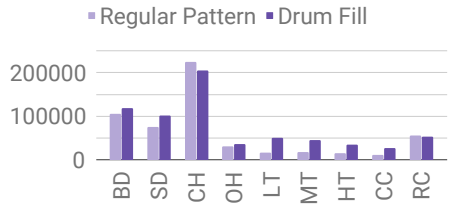


Fig. 2: Amount of notes by instrument for the ML dataset

Proximity of Drum Fills We want our extracted drum fills not too close to each other, as we consider that fills only appear every 4, 8, or 16 bars. Thus, we compute the length of the longest serie of adjacent fills in every song of our dataset. The average lenght is 0.68 and 1.90 for the ML dataset and the RB dataset respectively. The perfect result would be 1, so it is close to what we expect.

Distribution of Genres We compute the average amount of fills extracted over genres. We expect to find more fills in the following genres: Metal and Jazz. Our RB dataset follows well this intuition but this is not the case for our ML dataset.

7 Generation of Drum Fills

Our main goal is to generate a bar containing a drum fill, conditioned by a previous bar containing a regular pattern. We decide to use an architecture often found in the Natural Language Processing state-of-the-art, many-to-many Recurrent Neural Networks (RNN), whose architecture is described in Figure 3.

7.1 Training

We train our model over 300 epochs with a batch size of 4096 for the RB dataset and 256 for the ML dataset. We use Adam [4] as optimization algorithm with a learning rate of 0.001 and binary cross-entropy as loss function. We remove from each dataset the intersection of the two datasets which we use later as a test dataset, in order to evaluate the model trained on different datasets. We use a split of 80/20 for the training/validation datasets.

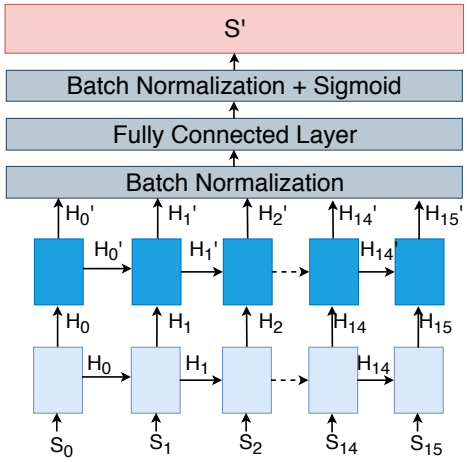


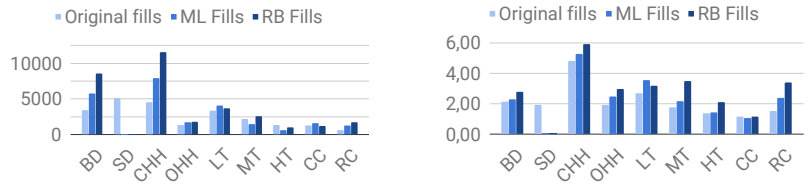
Fig. 3: Our model architecture is composed by two stacked GRU layers followed by batch norm and fully connected layers

7.2 Evaluation

We use the test dataset to generate two set of fills with our model trained on the ML dataset and the RB dataset. Then, we compare the original fills (ground truth fills) from our test dataset with the two other sets of generated fills (ML fills and RB fills)

Total of Notes by Instrument Applying the Pearson's chi-squared test between the total of notes by instrument of the three datasets (pair-wise), the p-value is less than 0.01, that shows that the fills are different in the three sets.

As main differences, we can see that the RB fills includes more bass drum and closed hit-hat than the other sets of fills. The ML fills include more low tom notes than the other set of fills as well. Unfortunately, our datasets, substracted from their high-density snares fills, do not allow us to create drum fills with snare rolls.



(a) Total (b) Standard-deviation

Fig. 4: Total and standard-deviation of the amount of notes in the generated/original fills by instrument

Diversity of Fills We take each set of fills, and we encoded them in the VAE's latent space of the Section 4.1. So, for each set of fills, we compute the sum of

the Euclidean distance between fills (pair-wise) as a measure of diversity for each set of fills. And give us 93012, 93844 and 102135 for the ML fills, the RB fills and the original fills respectively. We also compute the standard deviation of the amount of notes by instrument . From these two perspectives, we can see that the RB fills are more diverse than the ML fills but less diverse than the original fills.

8 User Study

Finally, we conduct a user study involving 51 participants (66% of musicians) recruited from the Internet. After a small test to know if participants are able to recognize a fill from a regular pattern (two people did not pass the test), people were asked to listen to 4 pieces of drums including a regular pattern repeated three times and then a drum fill. Two pieces of drums come from the ML fills and the RB fills respectively, the two other are the original fill and a "Rule based composed fill" (RC fill). The RC fill is the original regular pattern with the same toms/crash pattern added each time. We report in Table 3 the results. We can see that our methods do not beat a human for the task of composing a drum fill, even when the fill is composed with the same rule. Nevertheless, the RB fills are getting closer from the original fills. The original fill is not rated with a good grade in our experiment. In other words, human listeners do not think the human composed fills are good enough. Additionally, the RC fill has almost the same grade as the original fill. This indicates the subjective nature of the task and answers to the research question 3.

	ML	RB	Original	RC
Overall grade	2.61	2.90	**3.13**	3.10
Most coherent	17%	18%	29%	**36%**
Less coherent	**30%**	**30%**	23%	18%
Best groove	13%	25%	**34%**	28%
Worst groove	**35%**	30%	18%	17%

Table 3: Results of the user study, averaged over 49 subjects. The mean of the five-point scale grade is given in the first line. For the rest of the lines, the ratio of vote is given.

9 Conclusion

We have presented several axes to research in the field of drum fills detection and generation. We have shown the importance of considering a fill as a variation rather than through an absolute view. The results of our generation pipeline (detect drum fills with a rule-based method and then generate them with an RNN) are getting closer from the human-composed fills. In future work, we will explore fusion-method that combines machine-learning and rule-based method to improve the results of the drum fills detection with the help of more hand labelled data.

References

1. Raffel,C.: Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching. PhD Thesis (2016)
2. Dong, H.W.,Hsiao, W.Y., Yang, L.C., Yang, Y.H.: MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In:Thirty-Second AAAI Conference on Artificial Intelligence (2018)
3. Roberts, A., Raffel,C. , Engel,J., Hawthorne, C., Eck, D.: A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In: ICML 2018 (2018).
4. P. Kingma, D., Ba,J.: Adam: A method for stochastic optimization. In: ICLR 2015 (2015).
5. López-Serrano, P., Dittmar, C., Müller, M.: Finding drum breaks in digital music recordings. In: International Symposium on Computer Music Multidisciplinary Research (2017)
6. Lehner, B., Widmer, G., Sonnleitner, R.: On the reduction of false positives in singing voice detection. In: (ICASSP) (pp. 7480-7484). IEEE. (2014)
7. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Layton, R. (2013). API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning ,pp 108–122 (2013)
8. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: The International Conference on Learning Representations (2014)
9. Dong, H. W., Yang, Y. H.:. Convolutional generative adversarial networks with binary neurons for polyphonic music generation. In: ISMIR (2018)
10. Schroedl, S.: Play Drums Today. Hal Leonard. ISBN 0-634-02185-0 (2001)

Discriminative Feature Enhancement by Supervised Learning for Cover Song Identification

Zhesong Yu, Xiaoou Chen, and Deshun Yang

Institute of Computer Science and Technology, Peking University
{yzs, chenxiaoou, yangdeshun}@pku.edu.cn

Abstract. Cover song identification is to identify, for a query audio track, other recordings of the same composition, which shows potential use in music management and license protection. Existing methods involve complex sequence matching algorithms and hand-crafted features, where a further breakthrough is hard to be achieved. In this paper, exploiting large-scale data, we explore several supervised learning methods to enhance hand-crafted features and extract discriminative features for cover song retrieval. Experimental results show that the enhanced feature highly improve the precision on several datasets with low time complexity.

Keywords: Cover Song Identification, Supervised Learning, Enhancement

1 Introduction

With the growth of music collections on the Internet, how to manage and organize such extensive resources becomes a challenging task. Cover song identification, which is defined as retrieving the covers of a given song in a dataset, could be used to deal with these problems. Identifying cover songs can help detect copyright infringements and cope with music license management. Owing to its potential applications, it has been extensively studied over the past years. However, since variations of timbre, tonality and music structure usually exist in cover songs, it remains a challenging problem until now.

To address these challenges, researchers apply sequence alignment algorithms to this task. Chroma [6] extracted from music is used to represent the melody of music. Then, matching methods are utilized to measure the similarity among the chroma sequences. For instance, Ellis and Poliner cross-correlate beat-aligned chroma to find optimal matches of cover songs [5]. Dynamic Time Warp (DTW), a well-developed algorithm, is explored for sequence matching [7, 15]. Basic Local Alignment Search Tool for biosequence alignment is adapted to this task [12]. Serra et al. utilize non-linear time series analysis and develop a Dynamic Programming algorithm Qmax [17]. These approaches achieve high precision on small-scale datasets, typically containing hundreds of songs.

Nevertheless, these methods include complex sequence matching procedures, which lead to high time cost and are unsuitable for larger datasets. Some researchers devise compact representations from chroma instead of alignment algorithms. For example, Serra et al. model chroma sequences with time series [16]. 2D Fourier Magnitude (2DFM) is extracted from beat-aligned chroma to measure the similarities of songs [3]. In [1, 9, 2], cognition-inspired and chord descriptors are used to represent music. However, high generalizations of feature sequences lose tremendous temporal information and result in worse accuracy compared to alignment methods.

Moreover, the aforementioned methods involve complex hand-crafted features and matching procedures. Alternatively, some researchers attempt to utilize deep learning. Qi et al. [13] employ triplet networks for feature learning towards cover song identification but does not obtain very good results. Chang et al. [4] exploit Convolutional Neural Network to measure the similarities of the songs given the cross-similarity matrix. This method could be viewed as a post-process for sequence alignment methods; thus it is hard to be applied to large datasets. Moreover, the networks are trained on small-scale datasets which are not public; it's hard to compare with them directly. To deal with this problem, we use the cover songs dataset *Second Hands Song 100K (SHS100K)* consisting of 100K tracks [19], which is the largest public dataset for cover song identification to the best of our knowledge.

In this paper, we exploit supervised learning to improve the performance of hand-crafted features. Classifier network, triplet network and LDA are respectively explored to achieve the discriminative enhancement. Experimental results on several datasets show that our approach achieves better performance even compared to sequence alignment methods with low time complexity. Our contribution can be summarized in two aspects: Supervised learning is utilized to improve the performance of cover identification, different from previous methods which focus on alignment algorithm and feature engineering. Besides, different supervised methods are empirically explored to further improve precision.

2 APPROACH

As shown in Figure 1, our method first extracts M-dimensional 2DFM [3] from audio, and then obtains M^*-dimensional features by supervised learning. The new representation is used to describe audio and measure the distances to references in the dataset. Then, a ranking list is returned as a result. Note that the representations of references are precomputed when the dataset is built.

2.1 Feature Extraction

As shown in Figure 1, Harmonic Pitch Class Profile, an enhanced chroma representing the intensity of twelve pitch classes within a frame, is extracted from audio following the settings of [17]. The output is a sequence of 12-dimensional vectors—2 vectors per second, each vector corresponding to 2100ms. Then, 2D

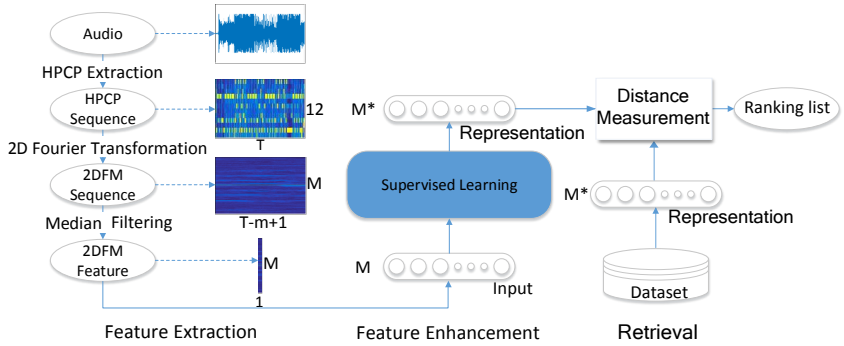
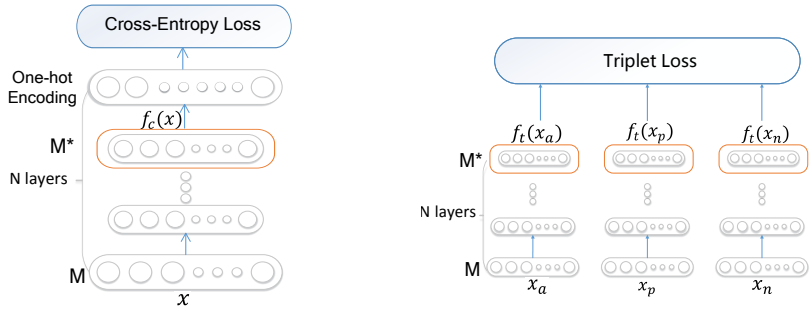


Fig. 1. Overview of the approach. The 2DFM is obtain from feature extration. Supervised learning methods transforms the feature from M dimensions to M^* dimensions.

Fourier Transformation is applied to m adjacent vectors with a hop size of 1, and the resulting feature has a shape of $M \times (T-m+1)$ (the magnitudes are reshaped into vectors, and M is equal to $12 \times m$, and T is the length of the HPCP). Finally, a median operation is applied to the feature along the row; thus we embed music into an M-dimensional vector. Additionally, Principle Component Analysis (PCA) is used to reduce the dimension further [17]. The prominence of this method lies in two points. First, it is robust against semitone rotations (key transpositions), which are common changes existing on cover songs. Second, the representation is compact and fixed-dimensional, which could be used to measure the similarity of two songs easily. In the experiments, m is set to 75 and M is 900.

2.2 Feature Enhancement

Indeed, 2DFM is well designed to reduce the impacts of variations in cover songs. However, scrutiny reveals some problems. First, with 2D Fourier transformations, the magnitude usually gets smaller for high frequency but get larger for low frequency. In other words, the magnitudes range over different scales; the similarities of 2DFMs measured by Euclidean distance will be governed by the large values, which undermine the precision of retrieval. Second, although PCA is utilized to extract discriminative and compact feature in [3], there is still room for improvement by supervised learning. Our solution is to utilize supervised learning to improve the performance of 2DFM. Important, the readers should note that in our task, we learn a discriminative feature based on a hand-crafted feature rather than low-level features. Hence, our approach should be viewed as techniques to enhance hand-crafted features rather than exploit deep networks for feature learning. Two kinds of neural networks are explored in our experiments, and LDA is provided as a comparison.



(a) Classifier Network (b) Triplet Network

Fig. 2. Network structure. Red rectangle denotes the enhanced feature.

Classifier Network Motivated by the success of adopting classifier networks for image retrieval [10], we train classifiers for cover song identification. Different covers of a song are viewed as the samples from a class, and different songs are regarded as different classes. As shown in Figure 2a, 2DFM is fed into the fully-connected networks to infer the one-hot encoding vectors representing the category of songs. After the learning, the last hidden layer is used to generate a new representation for retrieval. The number of layers N and neurons M^* is selected from $\{1, 2, 3, 4, 5\}$ and $\{300, 500, 1000\}$ respectively, and different activations like sigmoid, relu, tanh are empirically explored in the experiments. Cross-entropy and Adam optimizer are used to tune the network with a learning rate of 0.001.

Triplet Network Even though classifier networks optimize cross-entropy error, essentially they do not ensure that the learned representation performs well on retrieval. Alternatively, we use triplet networks, which minimize the distances between similar pairs and maximize the distances between dissimilar pairs.

As shown in Figure 2b, the network consists of three sub-nets sharing the same parameters. A triplet input is denoted as (x_a, x_p, x_n), where x_a indicates the anchor song and x_p measn the song of the same class while x_n denotes the song from a different class. The triplet is fed into the network and mapped into the feature space $F = f(x)$, where $f(\cdot)$ indicates the network function, and the new triplet is denoted as (F_a, F_p, F_n). The loss function is defined as:

$$\mathcal{L} = max(D(F_a, F_p) - D(F_a, F_n) + margin, 0) \tag{1}$$

where D represents cosine distance, and $margin$ is a hyperparameter. To minimize the loss function, learn a representation pushing D to 0 and D to be larger than $margin$. To train a more robust network, we follow a similar sampling scheme as [8]. One could review it for more detail.

We design our own strategy on triplet mining based on [8]. Firstly, P different categories are randomly picked from training set, and K versions for each

category are selected. Then their 2DFM features are fed to the net, and the cosine distances of each two audio embeddings are calculated. To construct a triplet (x_a, x_p, x_n), for an anchor x_a, the k_1-hardest positive is selected as x_p, meaning that the distance $D(f(x_a), f(x_p))$ is the k_1-largest distance between x_a and other samples of the same class; and the k_2-hardest negative is selected as x_n, meaning that the distance $D(f(x_a), f(x_n))$ is the k_2-smallest distance between x_a and other samples of the different class. In this way, it produces PK triplets, and the selected triplets are valid for training.

Linear Discriminant Analysis Besides the neural network-based algorithms, we utilize a conventional supervised method. Linear Discriminant Analysis (LDA) is a supervised approach for classification and feature reduction. It seeks new orthogonal axes maximizing inter-class variance and minimizing intra-class variance. In other words, the similar samples would be arranged closely and the opposite hold for the dissimilar samples. Note that the technique is used in retrieval tasks to improve precision. In our experiments, we iterate the hyper-parameter of LDA M^* through $\{200, 500, 900\}$ and tune it on the validation set.

2.3 Retrieval

After the training, the network is used to extract representations for retrieval. Cosine similarity is used to measure the similarity given the representations of two songs as follows:

$$s = 1 - \frac{u^T v}{|u||v|} \tag{2}$$

where s represents the similarity between the two representations u, v extracted from the network. Given a query, the retrieval system generates the enhanced feature from 2DFM and then computes the similarities to the references in the dataset and returns a ranking list.

3 EXPERIMENT AND DISCUSSION

3.1 Datasets

Second Hand Songs 100K (SHS100K) is collected from Second Hand Songs website [19]. *SHS100K* contains 108869 tracks with 9202 songs; each song has a different number of versions. In our experiments, we split this dataset into three sets—*SHS100K-TRAIN*, *SHS100K-VAL* and *SHS100K-TEST* with a ratio of 8 : 1 : 1.

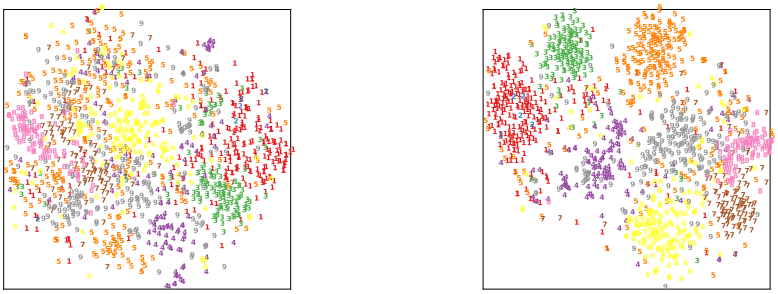
Youtube contains 50 compositions from multiple genres like rock, jazz and pop music, and each song has 7 versions with 2 original versions and 5 different versions. *Youtube* is used for testing, and the original versions are used as query and the others are used as references, the same as the setting of [19, 18].

Covers2473 constructed by ourselves contains 165 songs with 2473 recordings and is used for testing. It is composed of Chinese, Japanese, and Korean pop music. Since our network is trained on *SHS100K-TRAIN* mainly containing western music, we want to analyze the generalization ability of our approach on *Covers2473*.

3.2 Evaluation Metric

Mean average precision (MAP), precision at 10 (P@10) and mean rank of first correctly identified cover (MR1) are calculated to assess the performances. MAP is the mean of average precision, and P@10 is the mean ratio of covers identified in top 10 (note that it may refer to the mean number of covers identified in top 10 too). These metrics are the ones used in Mirex Audio Cover Song Identification contest [1]. Additionally, the large value is preferable for MAP and P@10 and the opposite holds for MR1. The query time is the time that a method spends retrieving the cover songs of a query, which doesn't include the time for feature extraction.

3.3 Representation Enhancement



(a) 2DFM (b) Improved Feature

Fig. 3. T-SNE embedding of cover songs. (a) is the embedding of 2DFM. (b) is the embedding of improved features produced by Feed-Forward Networks.

Firstly, we train our models with different hyperparameter settings on *SHS100K-TRAIN* and tune on *SHS100K-VAL*. LDA obtains its best performance with $M^* = 200$, and the networks obtain similar precisions under different settings. However, we observe that with shallow layers and linear activations, the networks have higher precision. This result might not be surprising since 2DFM

[1] https://www.music-ir.org/mirex/wiki/2017:Audio_Cover_Song_Identification

is a high-level descriptor. Moreover, during the training, the networks obtain very low training loss, meaning that training data is approximately linear separable. Hence, utilizing complex network structures may cause overfitting and undermine performance. Finally, we use a three-layer classifier network with 500 neurons and a three-layer triplet network with 300 neurons for comparison (see 3.4).

Furthermore, we visualize the representation and show insights on the reason why the enhanced feature improves the performance. Several songs with various renditions are randomly selected from the training set. Since 2DFM and embedding features are high-dimensional data, t-SNE [11](an algorithm to visualize high-dimensional data) is used to display them in Figure 3. Indeed, one could find that the intra-class variance becomes smaller, and the opposite holds for the inter-class variance. Ambiguity in decision boundaries on 2DFM is partially eliminated by the embedding space; thus it results in better retrieval results.

Table 1. Performances on different datasets (- denotes that the results are not shown in original works)

	MAP	P@10	MR1	Time
Results on *Youtube*				
DTW	0.476	0.124	11.2	4.93s
DPLA [15]	0.525	0.132	9.43	2420s
SiMPle [18]	0.591	0.140	7.91	18.7s
Fingerprinting [14]	0.648	0.145	8.27	-
KiCNN [19]	0.660	**0.158**	5.60	50.0us
2DFM [3]	0.448	0.117	12.2	88.9us
2DFM+LDA	0.628	0.147	6.93	40.0us
2DFM+Classifier Network	0.675	**0.158**	5.67	50.9us
2DFM+Triplet Network	**0.682**	0.155	**5.58**	**32.3us**
Results on *SHS100K-TEST*				
DTW	0.148	0.149	481	355s
2DFM[3]	0.104	0.113	415	12.6ms
2DFM+LDA	0.190	0.188	214	5.11ms
2DFM+Classifier Network	**0.226**	**0.208**	186	5.72ms
2DFM+Triplet Network	0.215	0.202	**174**	**3.83ms**
Results on *Covers2473*				
DTW	0.338	**0.452**	20.0	109s
2DFM[3]	0.199	0.272	35.5	2.52ms
2DFM+LDA	0.332	0.438	**13.0**	0.68ms
2DFM+Classifier Network	**0.347**	0.445	15.8	1.46ms
2DFM+Triplet Network	0.334	0.428	14.4	**0.67ms**

3.4 Comparison

2DFM and an alignment method based on DTW are implemented for comparison. For Youtube, we follow the same experimental setting of recent methods

(see Section 3.1) and report their results on *Youtube* instead of reproducing the methods. For *SHS100K-TEST* and *Covers2473*, we compute the similarities among all recordings for evaluation. Our approach is implemented in Python and C++ and run on a Dell PowerEdge R730 server with an Intel Xeon E5-2640v3 processor and two TITAN X GPUs.

As shown in Table 1, exploiting supervised learning highly improves the performance of 2DFM and attains better results on the three datasets consistently. Especially, the triplet network outperforms the state-of-the-art method on Youtube. Note that expect our approach and [3, 19], the others are sequence alignment approaches, relying on sophisticated matching algorithms. Besides, the improvement of *Covers2473* shows that the model does not only work well on western music but also has a good generalization ability.

Importantly, our method has a low time complexity as it avoids complicated matching procedure and extracts more low-dimensional representation than 2DFM. Thus, even though it requires extra time for feature extraction, the low-dimensional representation compensates the extra time and results in less query time. Note that [15, 18, 14] are implemented in operation systems and programming languages different from us. Hence, their query time is provided as a reference. Nonetheless, these methods are sequence alignment methods, having a similar complexity to DTW; DTW is used for comparison on behalf of them.

In practice, though the time complexity of our approach is proportional to the scale of the dataset, it is still acceptable to large-scale datasets. By estimation, it could process a query within 1s on the datasets with 1M songs. Moreover, the method could be used as a filtering scheme and combined with alignment methods to build practical cover song retrieval systems.

4 Conclusion

In this paper, we exploit supervised learning to enhance feature for cover song identification. Minimizing classification error or triplet loss, our approach reduces the distances among different versions of a song and learns a discriminative feature for cover song retrieval. Experimental results show the approach highly improves the performance of 2DFM and outperforms DTW and even alignment methods on three datasets. Owing to the lower dimension compared to 2DFM, it works effectively and could be used as a filtering scheme for large-scale datasets.

References

1. van Balen, J., Bountouridis, D., Wiering, F., Veltkamp, R.C.: Cognition-inspired descriptors for scalable cover song retrieval. In: proceedings of the 15th international conference on Music Information Retrieval (2014)
2. Bertin-Mahieux, T., Ellis, D.P.: Large-scale cover song recognition using hashed chroma landmarks. In: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 117–120. IEEE (2011)

3. Bertin-Mahieux, T., Ellis, D.P.: Large-scale cover song recognition using the 2d fourier transform magnitude. In: International Society for Music Information Retrieval Conference (2012)
4. Chang, S., Lee, J., Choe, S.K., Lee, K.: Audio cover song identification using convolutional neural network. In: Workshop Machine Learning for Audio Signal Processing at NIPS (ML4Audio@NIPS17)
5. Ellis, D.P., Poliner, G.E.: Identifying cover songs with chroma features and dynamic programming beat tracking. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2007)
6. Fujishima, T.: Realtime chord recognition of musical sound: a system using common lisp music. In: ICMC. pp. 464–467 (1999)
7. Gómez, E., Herrera, P.: The song remains the same: identifying versions of the same piece using tonal descriptors. In: International Society for Music Information Retrieval Conference. pp. 180–185 (2006)
8. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification (2017)
9. Khadkevich, M., Omologo, M.: Large-scale cover song identification using chord profiles. In: ISMIR. pp. 233–238 (2013)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
11. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
12. Martin, B., Brown, D.G., Hanna, P., Ferraro, P.: Blast for audio sequences alignment: a fast scalable cover identification. In: International Society for Music Information Retrieval Conference (2012)
13. Qi, X., Yang, D., Chen, X.: Audio feature learning with triplet-based embedding network. In: AAAI. pp. 4979–4980 (2017)
14. Seetharaman, P., Rafii, Z.: Cover song identification with 2d fourier transform sequences. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 616–620 (2017)
15. Serra, J., Gómez, E., Herrera, P., Serra, X.: Chroma binary similarity and local alignment applied to cover song identification. IEEE Transactions on Audio, Speech, and Language Processing **16**(6), 1138–1151 (2008)
16. Serra, J., Kantz, H., Serra, X., Andrzejak, R.G.: Predictability of music descriptor time series and its application to cover song detection. IEEE Transactions on Audio, Speech, and Language Processing **20**(2), 514–525 (2012)
17. Serra, J., Serra, X., Andrzejak, R.G.: Cross recurrence quantification for cover song identification. New Journal of Physics **11**(9), 093017 (2009)
18. Silva, D.F., Yeh, C.C.M., Batista, G.E.d.A.P.A., Keogh, E., et al.: Simple: assessing music similarity using subsequences joins. In: International Society for Music Information Retrieval Conference (2016)
19. Xu, X., Chen, X., Yang, D.: Key-invariant convolutional neural network toward efficient cover song identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2018)

Perception of the object attributes for sound synthesis purposes

Antoine Bourachot, Khoubeib Kanzari, Mitsuko Aramaki, Sølvi Ystad and Richard Kronland-Martinet

Aix Marseille Univ, CNRS, PRISM, Marseille, France
bourachot@prism.cnrs.fr

Abstract. This paper presents a work in progress on the perception of the attributes of the shape of a resonant object. As part of the ecological approach to perception – assuming that a sound contains specific morphologies that convey perceptually relevant information responsible for its recognition, called invariants – the PRISM laboratory has developed an environmental sound synthesizer aiming to provide perceptual and intuitive controls for a non-expert user. Following a brief presentation of the different strategies for controlling the perceptual attributes of the object, we present an experiment conducted with calibrated sounds generated by a physically-informed synthesis model. This test focuses on the perception of the shape of the object, more particularly its width and thickness since these attributes, especially the thickness, have not been much studied in the literature from a perceptual point of view. The first results show that the perception of width is difficult for listeners, while the perception of thickness is much easier. This study allows us to validate the proposed control strategy. Further works are planned to better characterize the perceptual invariants relevant for shape perception.

Keywords: Perception of shape, sound synthesis, perceptual invariant, impact sounds, intuitive control

1 Introduction

This paper presents a study on the perception of the shape of an object through listening to the sounds produced by this object. It is a work in progress and takes part of the development of an environmental sound synthesizer (environmental sounds are defined by all naturals sounds other than speech and music [1]).

We know from various studies that it is possible to make links between certain physical parameters of the object and the acoustical parameters of the sounds emitted. For example Lakatos et al. [2] conducted a study on wooden and metal bars of the same length, but of different heights and widths. They showed that subjects were more efficient at judging the height-width ratio when the difference between the bars was large. Carello et al. [3] revealed through the sounds of wooden sticks that there was a strong correlation between perceived and actual size. Tucker et al. [4] studied the perception of the size and shape of different

plates (square, circular, triangular) in different materials. They showed that the shape was difficult to recognize from the impulsive sounds and that the recognition of the material was independent of the shape of the object.

The study presented in this paper is in line with the ecological approach to perception, as proposed by Gibson [5]. First developed as part of the vision, this approach was extended to the hearing by Warren & Verbrugge [6] and then formalized by McAdams & Bigand [7]. The ecological approach to perception stipulates that our perception is not only governed by the capture of "physical stimuli" but also by the environment around the listener. It proposes that our perception is based on invariant structures contained in the stimulus. These invariants are categorized into two groups: structural invariants characterizing the physical properties of a sounding object and transformational invariants describing the action exerted on this object.

Based on the work of W.W. Gaver [8] [9], Aramaki et al. [10] have proposed a paradigm called action-object in which a sound is described as the result of an action performed on an object. This paradigm has been developed to allow intuitive control of sound synthesis. Initially Aramaki et al. [10] focused their studies on the perception of the material in the context of impact sounds and proposed structural invariants describing the perception of materials. Conan et al. [11] identified different transformational invariants, extending actions possibilities to rolling, friction and scratching interactions. Thoret et al. [12] then showed the existence of transformational invariants related to the auditory perception of biological movements.

A synthesizer has been developed at the PRISM laboratory based on these previous studies and serves as an experimental and demonstration platform.

This paper is organized as follows: first, we describe the different control strategies developed in the synthesizer for controlling the perceptual attributes of the object. Then we will move on to a perceptual test conducted by using a physically-inspired model proposed by Pruvost et al. [13]. We considered physical model rather than real sounds to be able to generate parameterized and continuous morphings between different shapes of objects: from a string to a membrane, and to a cube. The goal of this test is to evaluate the perception of width and thickness of an object since these attributes have not been much studied in the literature. The results will be discussed with respect to the information that they provided. Then we will end with propositions for further works to isolate invariants specific to the perceived shape of objects.

2 Physically-Based Control Strategy

The synthesizer developed at PRISM is based on the notion of sound invarians and offers the user an intuitive control interface for sound design. Indeed, the high level control proposed to the user is based on a semantic description of the action and the object involved [14], as shown in figure 1.

The control strategy is based on 3 hierarchical levels allowing to move from the high-level control related to the evocations of sound source attributes to the

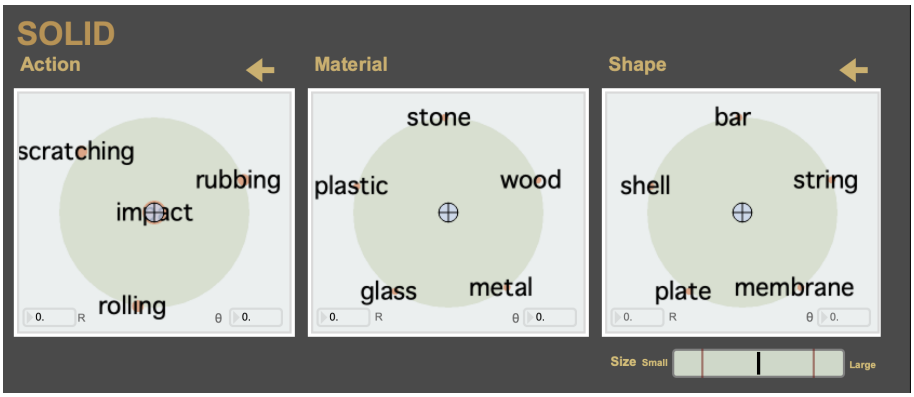


Fig. 1. Presentation of the high level controls of the synthesizer developed at PRISM

low-level of synthesis parameters. Between these two levels is the middle-level linked to the sound descriptors.

The semantic description of the perceived object is done by three control parameters. One for the material, one for the shape and one for the size. Concerning the material and shape controls, the user can navigate in a continuous space between the categories. Material control was mapped by Aramaki et al. [10] The design of the shape control was based on the analytical solutions for predefined shapes (string, bar, plate, membrane, shell) and linear interpolation was done between signal parameters (frequencies, amplitudes) to allow transitions from one label to another. A perceptual study was carried out to test the intuitive aspect of these synthesis' controls proposed. It emerged that the material space was satisfactory while the shape control was not very intuitive when people were asked to reproduce sounds created by the synthesizer. This result was expected since the shape space has not been formally calibrated yet from a perceptual point of view.

Following this study, Pruvost et al. [13] proposed a physics-inspired strategy for the control of the perceptual attributes of a sounding object. This control was for instance used in a framework of interactive and real-time synthesis of solids interaction, driven by a game engine. The controls proposed in this context were no longer semantic, but geometric (see figure 2). The object being defined by its length, width, thickness and so on, which is much more convenient for a control from a game engine.

Control of the Material: Aramaki et al. [10] confirmed previous studies showing that the modal damping plays a leading role in the auditory recognition of the materials. The authors proposed a damping model for the control of the perceived material that follows this equation:

$$\alpha(f) = \ e^{(\alpha_G + 2\pi f \alpha_R)} \tag{1}$$

The couple α_G and α_R, respectively the global and relative damping, was perceptually calibrated to characterize a given perceived material [10] .

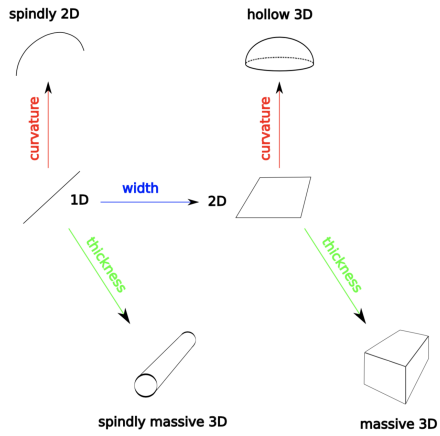


Fig. 2. Control parameters for the object based on its geometrical attributes (scheme taken from Pruvost et al. [13])

Control of the Shape: According to Rakovec et al. [15], the dimension of an object – one dimension of the object is considered when it is much larger than the others – and the massive/hollow aspect of the object are perceptually relevant parameters for the recognition of the shape.

In order to propose a continuous morphing between one-dimensional shape (e.g., a string) and a two-dimensional shape (e.g., a membrane), Pruvost et al. [13] defined that the modal distribution of sound evolves according to the following laws:

$$f_{nm}(\alpha, L) = f_0(L)\sqrt{\frac{m^2}{\alpha^2} + n^2} \tag{2}$$

With L the length (largest dimension) of the membrane. $\alpha \in]0, 1]$ so that the width $W = \alpha L$. This is reflected by an increase of spectral richness of the sound due to the emergence of numerous modal components according to α. We take into account a perceptual point of view when α tends toward 0: since m-related modes tend towards infinity and thus go beyond the limits of human hearing

we do not take into acount these modes and we consider that $f_{nm}(\alpha, L)$ tends toward $f_n(L)$:

$$f_{nm}(\alpha, L) \xrightarrow[\alpha \to 0]{} f_n(L) = f_0(L)n$$

With $f_n(L)$: the harmonic spectrum of the string.

To take into account the thickness of the object, the previous equation is extended to a third parameter, in analogy to the formula for the resonance frequencies of a rectangular cavity.

$$f_{nmp}(\alpha, \beta, L) = f_0(L)\sqrt{\frac{m^2}{\alpha^2} + \frac{p^2}{\beta^2} + n^2} \xrightarrow[\beta \to 0]{} f_{nm}(\alpha, L) \tag{3}$$

With $\beta \in]0, 1]$ the thickness $T = \beta L$ and $T \leq W \leq L$. In the same way, from a perceptual point of view, when $\beta \to 0$ we find the previous equation: (2). To our knowledge, there is no perceptual or physical study concerning the evolution of the damping with respect to the thickness of the object. However, some preliminary experiences led us to suggest that the damping should be modified while the thickness increases [13]. Therefore, we propose that the global damping α_G from equation (1) is increased as follows regarding the thickness linked to β:

$$\tilde{\alpha_G} = 2\beta\alpha_G \tag{4}$$

3 Perceptual Test

We therefore conducted a perceptual test to evaluate the control of shape provided by the previous model. In particular, the purpose of the test is to assess the perceived width and thickness of an object from the produced impacted sound.

3.1 Experimental Setup

Stimuli: In order to test this model, 36 sounds were selected from a "width-thickness" space that has been sampled regularly with a constant width step: $\alpha = 0; 0.2; 0.4; 0.6; 0.8; 1$, The same step was used made for the thickness β. All the possible combinations of α and β, this gives us a total of 36 sounds. The chosen action was an impact and the chosen material was metal to favor the perception of resonances of the object.

Participants: A total of 24 volunteers (14 males, 10 females) participated to the test. All subjects had an audiogram before the test. All the participants presented a normal hearing.

Apparatus: This test was conducted in a quiet room, isolated from ambient noise. Sounds were played with an Apple Mac Pro 6.1 with Seinnheiser HD650 headphones. All the answers of the participants were collected on an interface developed with Max/MSP.

Procedure: After a familiarization session where some typical sounds of the model were presented to the subjects, the formal test began. The sounds were presented only once and in a random order for each subject. For each sound, the subjects were then asked to assess the width and thickness of the object that produced the sound by using sliders. The scale of the sliders for each dimension was between 1 and 100. We chose to present sounds one at a time since we aimed at evaluating whether the width and thickness could be perceived in the absolute (i.e., without reference) or not.

After evaluating all the sounds, the participants were asked to answer a questionnaire at the end of the experiment that contained the following questions:

- "According to you, this test was: very difficult, difficult, average, easy or obvious"
- "What do you think of the dimensions you played on?"
- "Do you have any other dimensional suggestions for your perception of shapes?
- "Do you have any other comments?

3.2 Results

For each set of collected perceptual scores (width and thickness), we performed a repeated measures analysis of variance (ANOVA) using STATISTICA, on all variables with Thickness and Width as factors. A Tukey HSD test was then conducted to specify significant effects. An effect was considered as significant for p-value lower than 0.05.
Participants found the test difficult. Many of them reported that they found the thickness much easier to perceive than the width.

Perceived Width: The ANOVA showed a significant main effect of the Width ($F(5,115)=28.098$, $p < 0.001$). The post-hoc tests showed that the two first levels of perceived width differed from each other ($p = 0.05$) and with the four others levels ($p < 0.001$).

As shown by the figure 3, the analysis also revealed a Width by Thickness interaction. It firstly showed that the width was relatively well perceived for the first thickness level ($\beta = 0$). This can easily be explained by the fact that the model used is based on the frequency distribution of a membrane, i.e. an object generally perceived very thin. In particular, we noticed that the first width, which corresponds to a string in the model, was perfectly recognized. We also see a similar evolution for the second thickness level ($\beta = 0.2$), with a constant overestimation for the first values. However, for the other thickness levels ($\beta > 0.2$), the participants perceived the width in an almost constant way.

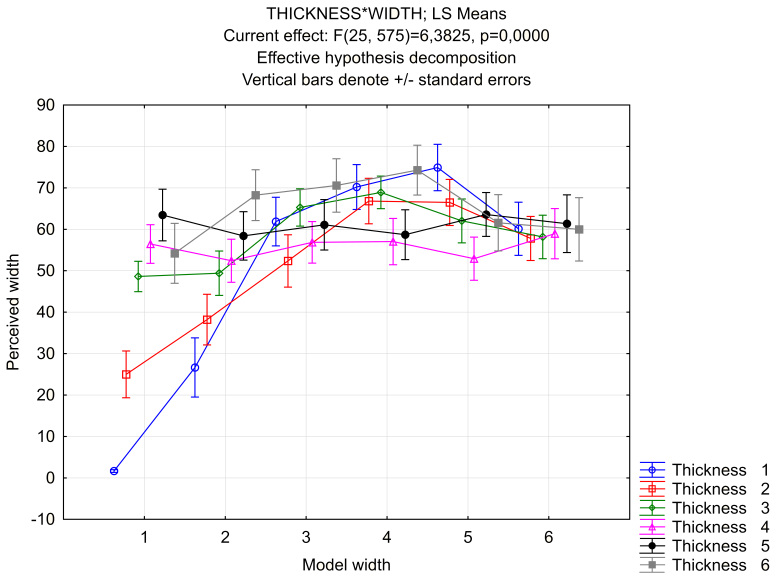


Fig. 3. Evolution of the perceived width regarding model width. Width 1, 2 ... 6 correspond respectively to $\alpha = 0, 0.2 \ldots 1$. Same for thickness with β

Perceived Thickness: The results are presented in the figure 4.

The perception of the thickness was really good for each level of thickness and width.

Each level of perceived thickness differed from each other significantly ($p < 0.02$): between the two first levels ($\beta = 0$ and 0.2) and between the others ($p < 0.003$).

These results suggest that the listeners were able to perceptually assess the thickness of an object. This consideration allowed us to validate the control strategy of the thickness involving a control of the damping (see equation (4)). This result is supported by self-report made by participants at the end of the experiment. Even if some of them report that it took time to understand the strategy used, none were disturbed by the link between thickness and damping.

4 Discussion and Further Works

The obtained results showed that participants did not perceive the width in an absolute way, except when the thickness is very small. By contrast, the thickness was well perceived for all width conditions. These considerations allow us to calibrate the control for the thickness from a perceptual point of view.

Future work will focus on the perception of width in a relative way, i.e., by comparison between sounds corresponding to different widths. We assume that

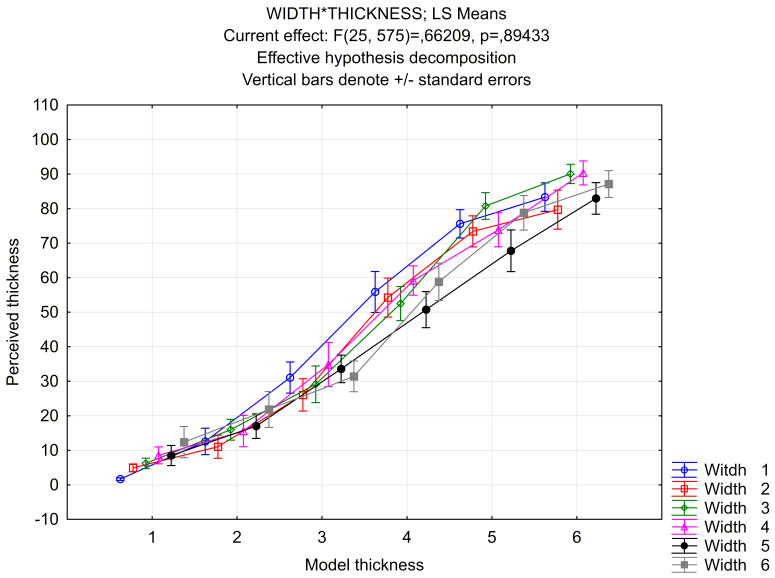


Fig. 4. Evolution of the perceived thickness regarding model thickness. Width 1, 2 ... 6 correspond respectively to $\alpha = 0, 0.2 \ldots 1$. Same for thickness with β

in this case the evaluation of width could be achieved easier. It will allow us to calibrate the width control in the synthesizer.

In summary, these studies will lead us to better characterize the relevant invariants specific to the perception of the object. To complete these studies, we aim at exploring numerical approach based on the morphology of the eigenmodes of objects. Indeed, to our knowledge, analytical physical models such as the ones used previously have limitations that do not allow us to reach all possible forms, for instance, a morphing between a thin plate and a cubic block. This part of the study is in progress.

The figure 5 shows the evolution of the modal distribution for thin plate - thick plate transition. Computing this evolution for different object sizes and geometries (circular, rectangular, etc.) will allow us to define a generic modal evolution law associated to the evocation of thickness, applicable to a signal synthesis model. Then these models will be assessed perceptually to calibrate the control strategy. Such models will be designed to have a small number of parameters in order to be able to offer intuitive control of the synthesis. The same process will be done for the other attributes of the object (curvature, hollowness, etc.).

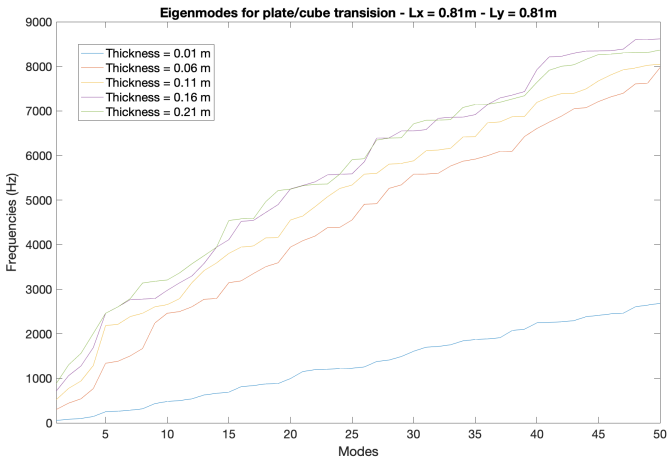


Fig. 5. Modal repartition for a thin plates (length = 0.81m, width = 0.81m, thickness = 0.01m) and different thick plate (constant thickness step: 0.05m)

5 Conclusion

In this paper we have made a brief review of the different synthesis controls proposed in the environmental sound synthesizer developed at the PRISM laboratory. We focused on the control of the shape of the sounding object.

The control of the shape of the object involved two main sound transformations: a 1D-2D string/membrane to bar/plate transition, and a 2D-3D plate/cube transition.

Then we presented a perceptual test related to this control, in particular the control of width and thickness, that had never been evaluated perceptually. In terms of perceptual expectation, the auditors had difficulty evaluating the width of the object, but had no problem evaluating its thickness. These results allowed us to validate the control of thickness in the synthesizer. Further works are underway to better define a control of the perceived width and more generally, to develop the search for the relevant invariants related to the perception of the shape.

References

1. B. Gygi, G. R. Kidd, and C. S. Watson, “Similarity and categorization of environmental sounds,” vol. 69, no. 6, pp. 839–855.
2. S. Lakatos, S. McAdams, and R. Causs, “The representation of auditory source characteristics: Simple geometric form,” vol. 59, no. 8, pp. 1180–1190.
3. C. Carello, K. L. Anderson, and A. J. Kunkler-Peck, “Perception of object length by sound,” vol. 9, no. 3, pp. 211–214.

4. S. Tucker and G. J. Brown, "Investigating the perception of the size, shape and material of damped and free vibrating plates."
5. J. Gibson, *The senses considered as perceptual systems.* Houghton Mifflin.
6. W. H. Warren and R. R. Verbrugge, "Auditory perception of breaking and bouncing events: a case study in ecological acoustics," vol. 10, no. 5, pp. 704–712.
7. S. McAdams and E. Bigand, *Thinking in sound: The cognitive psychology of human audition.*, ser. Oxford science publications. Clarendon Press/Oxford University Press.
8. W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," vol. 5, no. 1, pp. 1–29.
9. ——, "How do we hear in the world? explorations in ecological acoustics," vol. 5, no. 1, pp. 285–313.
10. M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, "Controlling the perceived material in an impact sound synthesizer," vol. 19, no. 2, pp. 301–314.
11. S. Conan, E. Thoret, M. Aramaki, O. Derrien, C. Gondre, S. Ystad, and R. Kronland-Martinet, "An intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling," vol. 38, no. 4, pp. 24–37.
12. E. Thoret, M. Aramaki, C. Gondre, R. Kronland-Martinet, and S. Ystad, "Controlling a non linear friction model for evocative sound synthesis applications," in *International Conference on Digital Audio Effects (DAFx)*, p. XX.
13. L. Pruvost, B. Scherrer, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Perception-based interactive sound synthesis of morphing solids' interactions," in *SIGGRAPH Asia 2015 Technical Briefs*. ACM.
14. M. Aramaki, R. Kronland-Martinet, and S. Ystad, "Perceptual control of environmental sound synthesis," in *Speech, Sound and Music Processing: Embracing Research in India*. Springer, pp. 172–186.
15. C.-E. Rakovec, M. Aramaki, and R. Kronland-Martinet, "Perception of material and shape of impacted everyday objects," pp. 943 – 959.

A Proposal of Emotion Evocative Sound Compositions for Therapeutic Purposes

Gabriela Salim Spagnol[1], Li Hui Ling[2], Li Min Li[1] and Jônatas Manzolli[3]

[1] School of Medical Sciences, University of Campinas, Brazil.
[2] Proactive Cultural Studio, Brazil
[3] Interdisciplinary Nucleus of Sound Communication (NICS, www.nics.unicamp.br)
gabrielaspagnol21@hotmail.com

Abstract. Recognition and understanding of emotions is a path for self healing. We have worked with Mandalas of Emotions, derived from the Traditional Chinese Medicine (TCM), as a complementary therapy. In this paper, we present the conceptual framework related to the creation of sound collages for the five elements of TCM and assessment of these compositions by experienced holistic therapists. Results present quantitative data, according to scales for relaxation, arousal and valence, and qualitative data from transcription and analysis of the recorded responses of volunteers. In our study, the most common perceptions were warmth, irritation, peace and fear. The innovation of this proposal may stimulate further research on emotion-evoking sounds, and in sound composition.

Keywords: Sounds, Music, Emotions, Integrative Therapies.

1 Introduction

Music and emotion has been long discussed, but systematic efforts to understand this relation are recent [1]. Its origin in the late 19th century occurs under a perspective of general psychology focused on psychophysics and experimental control. At this time, music psychology favored a more 'basic' perceptual and cognitive process related to music listening [2]. In the 1980s, Sloboda played an important role in developing the field of 'music cognition'. When his book, The Musical Mind [3], received recognition in the field, Sloboda had already started research in another field: music and emotion. Through a revival of Leonard B. Meyer's classic theory about musical expectations [4], Sloboda posed a correlation between 'cognition' and 'emotion'. Sloboda is now considered one of the driving forces in bringing 'music and emotion' to the spotlight, as a primary topic in music psychology [5].

Emotions, according to Koelstra et al. [6], are a psychophysiological process triggered by conscious and/or unconscious perception of an object or situation and are often associated with mood, personality, and motivation. Evoking emotions, including by means of music appreciation, is important to allow the recognition of feeling and to improve coping. In a healthcare service environment, sounds may be used as a masking tool, as a mean to improve patient-healthcare professional relation, and to

elucidate the emotional response to the current body and mind condition. This process may also mediate the creation of a therapeutic bound between patient and healthcare professional, the isolation of external sound interferences and improve patient experience and outcomes.

A specific dimensional approach for emotions, called the circumplex model of affection, proposes that all affective states result from two fundamental neurophysiological systems, one related to valence (a continuum of pleasure-dislike) and another to arousal or alertness [7]. According to the circumplex model, each emotion can be understood as a linear combination of two dimensions, or as varying degrees of valence and excitation. Joy, for example, is defined as an emotional state product of strong activation in the neural systems associated with valence or positive pleasure, together with the moderate activation in neural systems associated with excitation [8]. The affective states beyond joy also arise from the same two neurophysiological systems, but differ in the degree or extent of activation.

Specific emotions, therefore, arise from activation patterns within these two neurophysiological systems, along with cognitive interpretations and labeling of these central physiological experiences. Studies have applied the circumplex model to create and use musical parameters. In the study of Wassermann et al. [9], Sonification for Synthetic Emotions was used through the creation of an intelligent space, named as ADA. This artificial organism integrated a great number of sensorial modalities, so as to interact with the visitors through receptor systems. ADA used a language of sound and light to communicate their states of mind, emotions and behaviors.

We propose the use of circumplex model in a different context, in which we consider the five emotions based on Traditional Chinese Medicine (TCM), whose aims is to establishment a psychophysical balance. For this reason, the technique called Mandalas of Emotions (ME) applies nine steps to welcome emotions and develop abilities for reflection, as follows: identifying, accepting, accessing, revisiting, understanding, resignifying, reflecting, releasing emotions [10].

Fig.1. The cycle of mandalas in the sequence: spring/green, summer/red, high summer/yellow, autumn/white, winter/black.

For this process, this technique uses five colored, walnut-sized stones that are placed around the patient or on the person's abdomen for periods of 10 to 15 minutes, creating mandalas that correspond to five colors (green, red, yellow, white, black) and five emotions with its positive and negative correspondents (anger/comprehension, euphoria/ compassion, concern/gratitude, joy/ sadness, fear/courage) (Ling, 2013). These five colors establish a relation to the five seasons (spring, summer, high summer, fall, winter) and to the five functional systems (liver, heart, spleen and pancreas, lungs and kidney) [10], as depicted in Figure 1.

In this paper, we present the conceptual framework related to the creation of sound collages for the five Chinese elements (Wu Xing) and assessment of these compositions by holistic therapists. This relation was established by a strategy of sound collage, composing five pieces, one corresponding to each emotion.

Related work

Research has shown that sounds may translate emotions, as above mentioned, and also evoke emotions. In this sense, variations in sounds may elucidate what Huron [11] describes as the expectation-related emotion response system, which arouse corresponding limbic activations and contrasts.

Huron [11] defines five expectation-related emotion response systems: imagination (to motivate an organism to behave in ways that increase the likelihood of future beneficial outcomes), tension response (to prepare an organism for an impending event by tailoring arousal and attention to match the level of uncertainty and importance of an impending outcome), prediction response (to provide positive and negative inducements that encourage the formation of accurate expectations), reaction response (to address a possible worst-case situation by generating an immediate protective response), appraisal response (to provide positive and negative reinforcements related to the biological value of different final states). These concepts are applied in music composition in order to create absorbing sounds.

Moreover, psychologists describe the concept of entrainment as essential to perceive, react and enjoy music. Music, when considered as an external oscillator entraining a person's internal oscillators, potentially affects the sense of time and the sense of being in the world. Also, listeners exercise a great amount of their self-control in directing music entrainment, through unconscious processes and individual agency. Jones and colleagues published works between 1976 and 2002 on entrainment [12, 13]. This research considers three main assumptions on entrainment. First, human beings are considered inherently rhythmical, whose perception is capable of "tuning" with time patterns in the physical world. There is a tendency of synchronizing an individual's endogenous rhythms with perceived and expected rhythmic processes. Second, entrainment takes place as both period and phase present synchronization. At last, entrainment may vary in degrees.

In this paper, we illustrate the use of the Affective Slider[1] and other qualitative and quantitative data collection strategies performed to present sound compositions for expert assessment. Therefore, our methods section is intended to

[1] The Affective Slider, developed by Betella & Verschure [14] represents a model for data collection on reported valence and arousal, as described in the Methods Section. It is an advantage to use this model, since it will reflect a certain approximation to reality for experimental purposes.

support further studies with similar approaches. Results present quantitative data, according to scales for relaxation, arousal and valence, and qualitative data, which derive from transcription and analysis of the recorded responses of volunteers. In our study, the most common perceptions were warmth, irritation, peace and fear; drawing a parallel to its corresponding mandalas.

2 Methods

This section will present the method for creation of sound collages and its assessment by holistic therapists.

Creation of sound collages

We created five compositions, one for each mandala, using Audacity 2 with sound collages. These sound collages were chosen based in the elements of its corresponding mandala, which will be further elucidated in table 1. We also considered the stages of the relaxation process defined in the method, as three main phases (receive the emotion, reflect, and release the emotion). The duration of each composition was defined as 2.30 minutes, which creates a cycle that can be repeated for therapeutic purposes. Components followed the rationale described in Table 1.

Table 1. Elements, emotions, concepts and sound elements for each mandala.

Mandala	Elements	Emotions	Concepts	Sound elements
Green	Wood	Peace or Harmony	Beginning, birth of intentions	Shakuhachi and sound landscape of wind through bamboo
Red	Fire	Anxiety and confort	Growth, expansion	Sounds of hang drum, rattle and burning wood
Yellow	Earth	Gratitude and Concern	Harvest results	Sounds of two slowed hang drum rhythms
White	Metal	Joy and sadness	Reflection	Sounds of koshi bells and wind
Black	Water	Fear and courage	Conclusion	Sounds of ocean waves and rain

Original sounds have been altered from YouTube relaxation pieces to serve as samples and test the composition method. In order to create a database of sounds, we searched for videos on YouTube that presented a reference to the element of each mandala. For instance, wood is the element of the green mandala. The Japanese flute called Shakuhachi was chosen due to its relation to traditional knowledge and the possibility of having a melody with few notes, allowing collage-based techniques. We also used a sound of wind through bamboo to depict a sound landscape that posed a direct reference to wood and green, which the concept of mandalas (Table 1) associates with a calm and comprehensive atmosphere. For the red mandala, whose element is fire, there is a greater activity, with a rhythmic and repetitive melody, presented with a sound landscape of burning wood. This composition was expected to convey a feeling of warmth and a minor level of excitement or anxiety. For the yellow

mandala, we searched for sounds of traditional tribal drums as a reference to earth or desert, and we reduced the pace of the rhythm. For the white mandala, we searched for sounds created with metal, such as bells and chimes. At last, the black mandala comprises sounds related to water: sound waves, rain and a rain rattle.

The sound collages were organized to follow the nine stages of emotion recognition, as previously mentioned in the introduction section (identifying, accepting, accepting, accessing, revisiting, understanding, resignifying, reflecting and releasing emotions), which were comprised in three main steps: identify, revisit and release. Translated into sound composition, we created the rationale depicted in Figure 2 below.

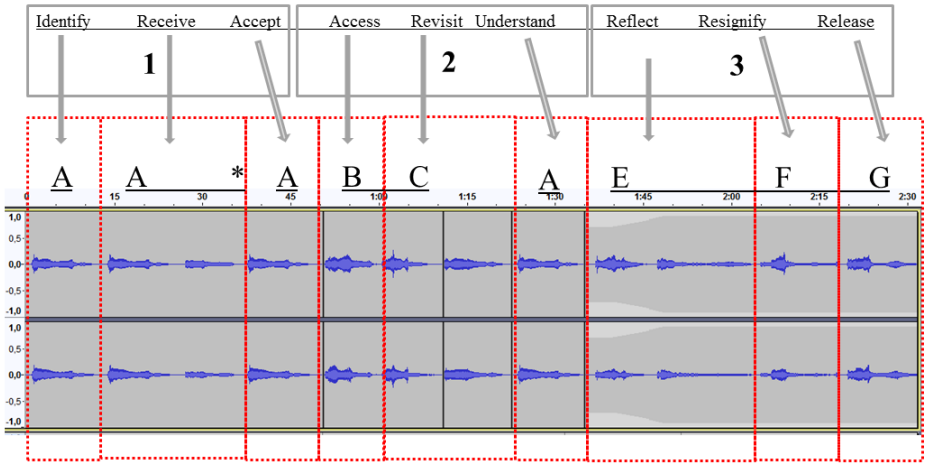


Fig. 2. The figure illustrates a print from the Audacity 2, in which A represents the arrival of the emotion, repeated throughout the composition.

In the next stage of this study, these collages will be reformulated, using original sounds and instruments, in order to be released for therapeutic purposes.

Assessment by holistic therapists

This study was performed with 8 participants on June 2018 from 8 am to 6 pm, with sessions of 30 minutes with each volunteer. In terms of sample size, we applied the first round of the Delphi method for validation of materials with experts as described by Alexandre and Colucci [15]. Selection process of volunteers included only professionals with more than one year of experience in applying Mandalas of Emotions, after signing an Informed Consent Term under Ethics Committee approval from University of Campinas. We prepared a controlled environment in which the volunteer laid down and listened to the compositions with a Microsoft Headset LX-3000. The volunteer listened to the 2.30 minutes samples in a randomized order, which was unknown to the researcher who performed the experiment.

Data collection started with questions regarding relaxation (visual Likert-scale, which ranges from 1, not relaxed, to 5, completely relaxed), valence and arousal levels with a validated visual scale called Affective Slider, that we printed in A4 white sheets to be marked with a pencil [14]. According to these authors, the "Affective Slider" (AS) is "a digital self-reporting tool composed of two sliders that measure arousal (top) and pleasure (bottom) on a continuous scale. The AS does not

require written instructions and it is intentionally displayed using a neutral chromatic palette to avoid bias in ratings due to the emotional connotation of colors".

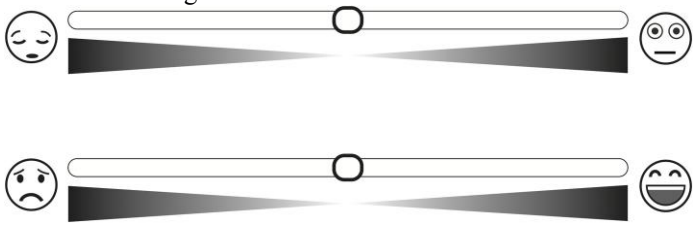


Fig. 3. This original Affective Slider is used in a touch-screen device, that allows to scroll the marker, placed in the center of each scale. We deleted this marker, printed this scale and instructed the participant to mark with a pencil.

Each mark on these scales was converted in centimeters, in order to elaborate graphs for further analysis. See the experiment flow in Fig. 4.

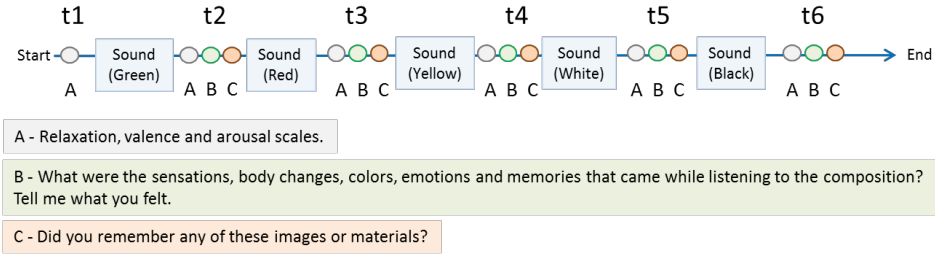


Fig. 4. Illustration of the experiment flow. The first period of data collection (t1) to the last (t6); A, B and C represent the qualitative and quantitative data collection questions and each box shows the presentation of sound. For each volunteer, the order of these compositions was different, in a randomized distribution. The sound was also unknown to the researcher who collected data.

Before listening to any composition, participants fulfilled the scales as mentioned in step A. In B, the researcher asked the volunteer response considering its body and mind perceptions, which include memories and colors. In C, the researcher presents a series of images (Fig.3) and materials (Fig.4) that represent each element (wood, fire, earth, metal and water) in forms that correspond to those used in the compositions. Those images and elements are depicted in figures 3 and 4. This strategy aimed to recall timbre perceptions, using these materials to enable characterizing the musical perception, once that volunteers did not present a background in musical knowledge.

Fig. 5. Pictures presented to volunteers that corresponded to elements of timbre in each composition.

Fig. 6. Materials presented to volunteers corresponded to elements of timbre in each composition.

Results

This section will present quantitative results, according to data collected using the scales for relaxation, arousal and valence, and qualitative data, which derives from transcription and analysis of the recorded responses of volunteers.

Quantitative results

Volunteers presented the level of relaxation in a continuous scale that ranged from 0 to 100 millimeters, marked with five possibilities of levels (1 = not at all, 5 = very much). Data in figure 5 represents measurement from the first period (t1) to the last (t6). Results show a progressive relaxation effect throughout the experiment, with a plateau between t5 and t6. It is important to mention that the compositions were in a randomized order; therefore, graph 1 depicts the isolated effect of the experiment, showing that any combination of compositions provokes a similar outcome throughout time.

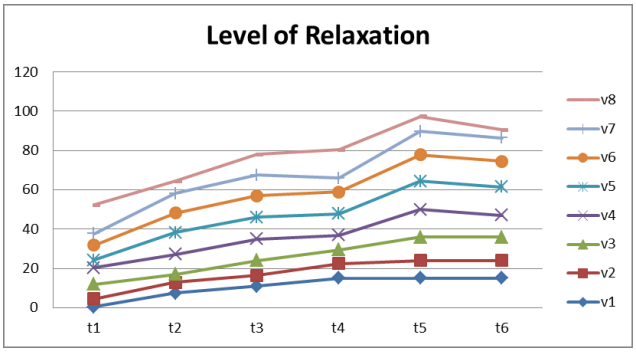


Fig. 5. Level of relaxation for each of the data collection periods, which range from t1 to t6. The initial period (t1) depicts the volunteer's baseline, before any sound intervention.

Volunteers were also required to fulfill the Affective Slider scale, which was printed in a paper, indicating with a pencil the current position of their arousal level, as shown in Fig. 6. The same was performed for valence (Fig.7). In these graphs, we present the isolated results per composition (green, red, yellow, white or black).

Considering media and standard deviation, results indicate a greater effect in terms of arousal variations in the following order: white, 0.7 (1.3); black, 0.3 (1.5); green, 0.4 (2.0); red 0.3 (1.5); yellow, 0.0 (2.0) and in terms of valence variations:

green, 2.35 (2.2); white 1.1 (2.0); red, 0.55 (0.75); black, 0.1 (2.1) and yellow, 0.0 (1.5). For therapeutical purposes, arousal levels must present a minor change, since the subject is expected to have a steady state of mind and body but, at the same time, present a variation related to the emotion-evoking process.

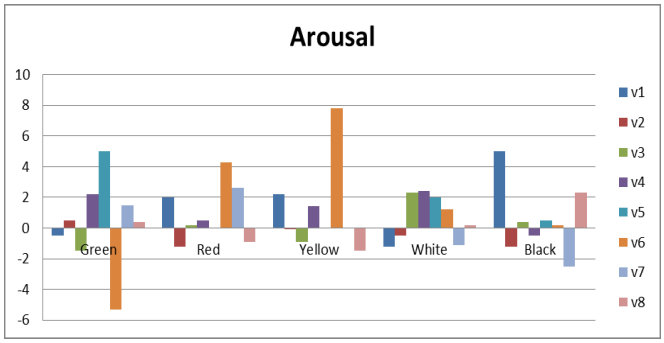


Fig. 6. Level of arousal for each mandala composition (green, red, yellow, white and black) and volunteers (v1 to v8).

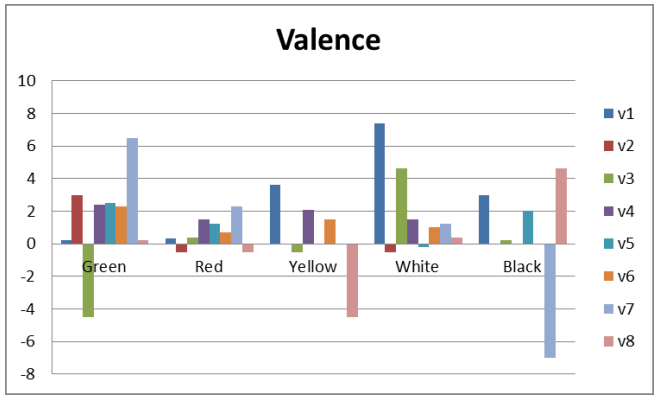


Fig. 7. Level of valence for each mandala composition (green, red, yellow, white and black) and volunteers (v1 to v8).

Qualitative results

For each period illustrated in the figure 1, that depicts the experiment flow, we asked questions that would convey the perceptions of our volunteers. For each composition, volunteers were asked to describe it as an emotion. We present this data in the form of a word cloud in figure 8, showing a wide range of outcomes.

Fig. 7. Results of the question "Which emotion arouse when listening to the sound composition?". Created with <https://www.wordclouds.com/>

Responses to questions B and C: "What were the sensations, body changes, colors, emotions and memories that came while listening to the compositions?" and "Did you remember any of these images or materials?", respectively, were analyzed for each composition in order to convey the main topics of outcomes, as described in the following paragraphs and in table 2.

In general, the composition of the green mandala brought a feeling of opening, beginning, with light tones. Six of eight volunteers pointed out as very relaxing, and that it should be the first of all compositions. According to volunteers, this sound referred to the green, wood, bamboo, as well as the desert and the vastness. These perceptions confirm that specialists related this composition to the green mandala. The sound of the red mandala was indicated by 7 of the 8 volunteers as related to warmth, comfort. Volunteers attributed this sensation to the cracking of fire sounds, but referred that its intensity could be reduced.

Specialists related the composition of yellow mandala to elements of earth (sand), wind, wood, and feelings of trust and gratitude. The sound of the white mandala brought elements of metal, water, wind, peace, but also irritation and anxiety associated with metallic sounds. Volunteers suggested reducing the information in this mandala. The sound of the black mandala was related to the sounds of waves, sand and also fear. Participants referred that there were several elements in this composition, and that these could be reduced.

Table 2. Responses for emotions, colors, elements and memories indicated by volunteers for each sound composition.

Mandala	Volunteers report
Green	Emotion: hope, peace, calm. Colors: light colors, green. Elements: bamboo, wood. Memories: forest, wind.
Red	Emotions: trust, gratitude, good memories. Colors: red, dark colors. Elements: fire. Memories: fire, bonfire.
Yellow	Emotions: peace, irritation, anxiety.

	Colors: light and warm colors. Elements: sand. Memories: wide field, horses.
White	Emotions: peace, irritation, anxiety. Colors: dark, blue, black. Elements: metal, water. Memories: wind, desert.
Black	Emotions: fear, irritation, affliction, determination. Colors: black, yellow, blue, dark. Elements: sand, water. Memories: storm, beach.

Discussion

This study presents the application of a method to assess specialists' response to collage-based compositions. These compositions are related to the five emotions derived from the Traditional Chinese Medicine (see table 1). Quantitative and qualitative data conveyed participants' perceptions and suggestions concerning the sounds, which will be later considered to compose the final version of each mandala.

The composition aesthetics followed characteristics of American Minimal Music and music commonly used for therapeutic purposes. According to a systematic review of randomized controlled trials that applied music interventions in a Neonatal Intensive Care Unit (NICU), music for therapies should be "soothing and not use too many different elements in terms of instruments, rhythms, timbres, melodies and harmonies" [16]. Considering this definition, the review study shows that the preferable choice of music is a lullaby, softly sung or played on an instrument. Also, we understand that familiarity plays an important role in music appreciation, so this strategy aimed at creating some degree of recognition related to a music style. In a study by Pereira et al. [17], brain activation data revealed that broad emotion-related limbic and paralimbic regions as well as the reward circuitry were significantly more active for familiar relative to unfamiliar music.

Instead of presenting a classification of emotions for participants to choose from, we performed open-ended questions, which were recorded and later analyzed. We also used physical elements and images related to elements of the five emotions to question whether participants identified sound landscapes and concepts applied in compositions. When comparing results in table 2 with emotions in table 1, we may state that there is a considerable parallel, and adjustments that can be implemented to reduce unwanted reactions, such as irritation.

Considering that these compositions allowed emotion arousal, as described in the Results section, we may compare our findings to those of Sloboda [18]: seventy-six college students were asked to indicate which of 25 emotions they had experienced to music. Sadness and joy were the two emotional states experienced by most listeners (96.2 and 93.4 percent, respectively). In our study, the most common perceptions were warmth, irritation, peace and fear as shown in Figure 2. We understand that the emotion of joy (55%) may be related to feelings of gratitude, welcoming, peace, comfort and calm, whereas sadness (37%) may be associated with fear, affliction, not to mention the report of "sadness" itself (See figure 7).

Limitations of this study are related to the sample size, which could be later expanded, and the application of only one round with experts. Once we apply changes

in the compositions, we expect to organize a second round of experts' assessment to validate our compositions. Our results are limited to the self-reported perceptions of volunteers, using visual scales that may not correspond to physiological changes in arousal and valence. Since this research project is under development, at some point, new research could be incorporated in this study. Also, quantitative and qualitative methods of conveying perception may be later complemented by physiological measures, such as heart rate and skin conductance, and, if applicable, brain activation experiments.

Conclusion

Mandalas of Emotions derive from a secular culture, Chinese Medicine, and establish a bridge between East and West for a need as ancient and complex as human beings: self-healing. Evoking, communicating and understanding emotions have been widely developed through integrative therapy and music, paths to reestablish a balance in the body and mind. The creation of sound compositions based in an emotion-evoking therapy may enhance its potential and, therefore, the possibility of self-healing.

This study provides a conceptual framework for creation of sound collages and testing of these with experts, based in Mandalas of Emotions. The innovation of this proposal may stimulate further research on emotion-evoking sounds, in sound composition and, possibly, in computational music. We understand that the creation of modulated music and systematization of sounds must be preceded by a process of applying and validating a conceptual framework, which could be pursued as we proposed. In the next stages of this ongoing project, we intend to use a music software tool to transform these compositions based on the listener's response in real time, and improve data collection methods.

References

1. Juslin, P. N., Sloboda, J. A.. The past, present, and future of music and emotion research. In: Juslin, P.N. & Sloboda J.A. (eds.). Handbook of music and emotion: Theory research, applications. pp. 933-955. New York, Oxford University Press (2010).
2. Deutsch, D. (ed.). The psychology of music (2nd ed.). New York, Academic Press (1999).
3. Sloboda, J.A. The Musical Mind. London, Clarendon Press (1986).
4. Meyer, L.B. Emotion and meaning in music. Chicago, IL, Chicago University Press (1956).
5. Thompson, W. F. Music, thought, and feeling. Understanding the psychology of music. Oxford: Oxford University Press (2009).
6. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., … Patras, I. Deap: a database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18-31 (2012).
7. Russell, J.A. A circumplex model of affect. Journal of Personality and Social Psychology 39, pp. 1161– 1178 (1980).
8. Posner, J., Russell, A., Bradley, J., Bradley, P.. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and

psychopathology. Development and psychopathology 17, pp. 715-34 (2005). 10.1017/S0954579405050340.
9. Wassermann, K.C., Eng., K., Paul, F.M.J., Manzolli, J. Live Soundscape Composition Based on Synthetic Emotions. IEEE Computer Society (2003). 1070-986X/03.
10. Ling, L.H. Dialogando com as emoções e promovendo a saúde. Curitiba, Insight (2013).
11. Huron, D. Sweet anticipation: music and the psychology of expectation. The MIT Press, New York (2006).
12. Jones, M.R.. Time, Our Lost Dimension: Toward a New Theory of Perception, Attention, and Memory. Psychological Review, 83(5), pp. 323-355(1976).
13. Jones, M.R. and Boltz, M. Dynamic Attending and Responses to Time. Psychological Review, 96(3), pp. 459-491 (1989). doi:10.1371/journal.pone.0027241
14. Betella, A., Verschure, P.F.M.J. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. PLoS ONE 11, p. e0148037 (2016). DOI: 10.1371/journal.pone.0148037
15. Alexandre, N.M.C., Coluci, M.Z.O. Content validity in the development and adaptation processes of measurement instruments. Ciência & Saúde Coletiva 16(7), pp. 3061–8 (2011).
16. van der Heijden MJE, Oliai Araghi S, Jeekel J, Reiss IKM, Hunink MGM, van Dijk M (2016) Do Hospitalized Premature Infants Benefit from Music Interventions? A Systematic Review of Randomized Controlled Trials. PLoS ONE 11(9), e0161848. doi:10.1371/journal.pone.0161848
17. Pereira, C.S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S.L. et al. Music and Emotions in the Brain: Familiarity Matters. PloS ONE 6(11), e27241 (2011). Doi: 10.1371/jornal.pone.0027241.
18. Sloboda, J. A. Empirical studies of emotional response to music. In M. Riess-Jones & S. Holleran (Eds.), Cognitive bases of musical communication. pp. 33-46. Washington, DC, American Psychological Association (1992).

Ackowledgements

We thank the support provided by the Interdisciplinary Nucleus for Sound Studies (NICS), UNICAMP, the Brazilian Research Institute of Neuroscience and Neurotechnology (BRAINN) and the Mandalas of Emotions experts. We also thank for the valuable insights from Charles de Paiva, researcher at NICS. This work was developed with the financial support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES), Financing Code 001.

On the Influence of Non-Linear Phenomena on Perceived Interactions in Percussive Instruments

Samuel Poirot[1], Stefan Bilbao[2], Sølvi Ystad[1], Mitsuko Aramaki[1] and Richard Kronland-Martinet[1]

[1] Aix Marseille Univ, CNRS, PRISM, Marseille, France
[2] Acoustics and Audio Group, James Clerk Maxwell Building, University of Edinburgh, EH9 3JZ, Edinburgh, United Kingdom
poirot@prism.cnrs.fr

Abstract. In this paper, we investigate the hypothesis that perceived impact strength is strongly influenced by the non-linear behavior produced by large deformations in percussive instruments. A sound corpus is first generated from a physical model that simulates non-linear vibrations of a thin plate. The effect of non-linear phenomena on the perceived strength is further quantified through a listening test. The aim of this study is to improve the expressive potential of synthesizers of percussive sounds through the development of signal transformation models. Future work will focus on the modeling of sound morphologies that correspond to non-linear behavior and the development of a transparent control strategy.

Keywords: Sound-Synthesis, Non-Linear Behavior, Auditory Perception, Impact Sounds, Control

1 Introduction

Digital sound synthesis enables the exploration of large sound spaces. One challenge is to find perceptually consistent control strategies to explore these spaces and hereby improve user expressivity.

For a synthesizer to be "expressive", it has been conjectured that the mapping between the inputs and outputs of the device must be transparent [1]. Transparency is here defined as an indicator of the psychophysiological distance, in the mind of the user and the audience, between the inputs and outputs of the instruments. Therefore, the control parameters and their link with the user's actions must be defined intuitively to allow for proper expression within a synthesizer. This has been investigated in several studies [2] [3] [4].

In the context of environmental sound synthesis, Conan proposed an intuitive control for continuous interaction sounds [5]. This work is a follow-up of the investigations of Aramaki et al. on the synthesis of percussive sounds with a control of the attributes of the impacted object (material, shape and size) [6][7][8][9]. These studies have led to the development of a sound synthesis environment with an intuitive control of the evoked action (impact, rolling, scratching, rubbing) and

object (material, size and shape) for solid interactions. Inspired by the ecological approach for auditory events [10][11][12], this environment has been developed within the action-object paradigm, which states that the sound is the result of an action on an object and that invariant structures can be independently associated to the actions and the objects. This synthesizer constitutes an experimental research environment for the PRISM laboratory that can be adapted for various uses, such as music practice, video games [13], or more generally sound design.

A dynamic mapping of the user's gesture with the synthesis parameters has been set up in specific cases to improve expressive control of the synthesizer, e.g. from squeaks to self-oscillating transitions for continuous interactions with a touch pad [14]. The perceived material has also been mapped with the strength of the impacts, and a drumstick has been instrumented to capture the gesture of the user. However, no work has so far been done to define the perceptual expectations related to the strength of the impact in a more general situation.

The objective of this study is to define expectations regarding the evolution of the sound radiated by an object when it is struck with different strengths. To this end, we study the effects of non-linear behavior on the radiated sound. The idea is to apply transformations to the signal for a transparent mapping of the strength of an impact for percussive sounds. In the context of virtual music instruments, this parameter could be linked to the velocity, a central parameter for expressiveness of percussive instruments on midi controllers.

Such an idea has been investigated in the musical context for years. Most virtual instruments offer dynamic behavior and adapt the sound rendering to the velocity of touch on the keyboard, especially for the highest values. A simple and convincing effect is to add brightness for high velocity values. For physical systems, this effect is due to the behavior of the impactor, whose deformation varies non-linearly with the pressure exerted [15]. For instance, the deformation of piano hammers follows a non-linear law related to their composition (wooden core surrounded by felt) [16][17]. Also, rules of thumb are proposed in various publications for a transparent mapping between the sound synthesis parameters and the user's gesture. One example is the mapping of the modulation index with the impact velocity in the context of FM [18], or the shapes of harmonic distortion laws to replicate a "physical" behavior [19]. In the context of environmental sound synthesis, one can also mention Warren and Verbrugge's investigations on sound morphologies responsible for breaking event recognition [20].

The proposed approach is to model the physical system and its non-linear behavior in order to synthesize realistic sounds for various configurations. The relevance of these phenomena with respect to perceived strength of the impact can then be evaluated through listening tests. Finally, we aim to model sound morphologies responsible for the evocation of high intensity impacts.

Physical modelling allows the generation of realistic sounds, and the mapping of the user's gesture is obvious and transparent since the synthesis parameters are physically meaningful. Moreover, the ever-growing computing capacities of commercially-available hardware makes it possible to consider real-time synthesis for increasingly complex models, opening the possibility of designing virtual

instruments with this type of synthesis [21][22][23]. However, the inherent constraints of a physical model limit the possibilities of sound creation. Indeed, the design of such a virtual instrument is based on modelling the behavior of an existing element rather than on the free description of a sound. The positioning of this study is to use physical models to generate a sound corpus that is further analyzed to extract and model the sound morphologies (signal model) corresponding to the phenomena studied. In this way, we free ourselves from the constraints of physical modeling by allowing the generation of metaphorical sounds. On the other hand, this type of approach, as presented by Rocchesso et al. [24], is complex because it implies three levels of research: auditory perception, physics-based sounds modeling, and expressive parametric control.

This paper presents a preliminary study for the development of perceptually controlled synthesis processes. We propose here to demonstrate that non-linear phenomena play a major role in the perception of the impact strength in the case of thin plates.

2 Review of Notable Non-Linear Behavior for Everyday Sounding Objects

Our perceptual expectations are driven by real sounding objects that we are used to hearing and manipulating. In this section, we briefly review the non-linear behaviors that can induce a notable transformation of the sound radiated by an object impacted with different strengths.

For this purpose we focus on percussive musical instruments, which sound timbre changes according to the excitation. The choice of the impactor (hammer, mallet, drum stick, brush, plectrum, fingers, hand palm) and the gesture (force profile, impact position) define the quantity of energy and its distribution on the resonator modes. In general, the impactor hardens as the intensity of the impact increases, inducing a brighter sound.

After excitation, the sound radiated by the instrument can be modified in case of interactions between the resonator and other elements. For example, the specific timbre of the tanpura and sitar is caused by the interaction of the strings with the bridge and the sympathetic strings. These interactions can lead to a muted sound (ghost notes, cymbal choke, palm mute), pitch modifications (natural and pinched harmonics, slide, bend, ghost notes, udu drum) and the appearance of other partial tones over time (slapping, string buzz, tanpura and sitar double bridge, snare wire, rattling elements in a cajon or a mbira).

Geometric non-linearities in the resonator itself may also occur for large amplitude deformations, resulting in time varying mode frequencies, harmonic distortion and mode coupling. These phenomena are particularly prevalent in the case of gong and cymbal crashing (wave turbulences). They can also be heard through timbre variations in steel drums, or pitch bending of strings and membranes due to large amplitude vibrations. Beyond such high amplitude vibrations, plastic deformations or ruptures can be observed.

These observations can be transposed to objects of everyday life. For example, the force with which someone knocks on the door may be recognized through the impactor (what part of the hand, and with which hardness), and the interaction with surrounding walls and rattling elements. Thin structures that admit large elastic deformations (e.g. metal sheets) behave similarly to cymbals and gongs. Weaker flexible structures deform (cans, plastic bottles), fibrous materials creak then crack (wood, composite), while more fragile structures crack then break (glass slides). Also, paper sheets, aluminum foil, plastic bags, fabrics, loose membranes and strings do not vibrate much due to their low stiffness. Their behavior is strongly non-linear, and the radiated sound is not tonal.

To sum up, the dynamic behavior of an impacted object may be due to the non-linear behavior of the object or non-linear interactions with other elements. A study has already been initiated to define a morphological invariant of non-linearity related to interactions [25]. In this paper, the objective is to check whether non-linear phenomena related to large vibrational amplitudes of an object are central to the perceived intensity of the impact.

3 Methods

We hypothesized that the perception of the strength of an impact is correlated with the occurrence of audible non-linear phenomena in the sound radiated by an object for large vibration amplitudes. It is assumed here that the strength of the impact is expressed by the object's ability to resist this impact. This implies that the perceived strength depends on the structural characteristics of the impacted object. To verify this hypothesis, an experiment was conducted in which subjects were to judge the strength of the impact and the object's ability to resist this impact by answering the following question: *In this test, we ask you to evaluate the strength of the impact and the "suffering" of the object for each sound* (In French, *Vous devrez évaluer la force de l'impact et la "souffrance" de l'objet pour chacun des sons*). The subjects were told that the "suffering" of an object reflects its difficulty to bear the deformation produced by the excitation. The degree of "suffering of the object" was used as an indicator for the perceived weakness of the object, supposed to be correlated with the occurrence of audible non-linear phenomena.

We chose to focus on the behavior of thin plates to test our hypothesis because it generates characteristic and easily recognizable sounds for large deformations.

This is a first approach to identify sound morphologies linked to audible non-linear phenomena, that can be used to define a signal transformation model that can be applied to different sound textures.

3.1 Synthesis of the Stimuli

Thin plate model. The stimuli were synthesized by numerical solving the Von Karman system, a widely used model of nonlinear vibration of plates at

moderate amplitudes with a quite compact form [22].

$$u_{tt} = -\frac{D}{\rho H}\Delta\Delta u + \frac{1}{\rho H}\mathcal{L}(\phi, u) - 2\sigma_0 u_t + 2\sigma_1 \Delta u_t + \frac{e}{\rho H}F(t) \tag{1a}$$

$$\Delta\Delta\phi = -\frac{EH}{2}\mathcal{L}(u, u) \tag{1b}$$

Where $u(x, y, t)$ is the transverse plate deflection, $\phi(x, y, t)$ is often referred to as the Airy stress function, $F(t)$ is the excitation force, $e = \delta(x - x_i, y - y_i)$ the 2D Dirac function, $\sigma_0 = 0.1$ rad$\cdot$ s^{-1} and $\sigma_1 = 0.001$m$^2\cdot$s^{-1} are the damping coefficients, H is the plate thickness, in m. The material is defined as steel, with a density of $\rho = 7800$ kg$\cdot$m^{-3}, D is set as a function of Young's modulus $E = 210$ GPa and Poisson's ratio $\nu = 0.3$ as

$$D = \frac{EH^3}{12(1 - \nu^2)}. \tag{2}$$

In Cartesian coordinates, the nonlinear operator $\mathcal{L}$ is defined by

$$\mathcal{L}(f, g) = \partial_x^2 f \partial_y^2 g + \partial_y^2 f \partial_x^2 g - 2\partial_x \partial_y f \partial_x \partial_y g \tag{3}$$

for any two arbitrary functions $f(x, y, t)$ and $g(x, y, t)$.

We set the length and width $L_x = 0.6$ m and $L_y = 0.7$ m, and we synthesized the sounds for three different thicknesses $H_1 = 1$ mm ; $H_2 = 2$ mm ; $H_3 = 3$ mm.

The boundary conditions were set free for $x = 0$ and $x = L_x$

$$u_{xx} + \nu u_{yy} = u_{xxx} + (2 - \nu)u_{xyy} = 0, \tag{4}$$

and simply supported for $y = 0$ and $y = L_y$

$$u = u_{yy} = 0. \tag{5}$$

These boundary conditions have been chosen to increase the occurrence of wave turbulences and limit frequency variations.

We chose to model the impact by a raised cosine rather than by a mallet model in order to be able to control the brightness independently of the strength of the impact.

$$F(t) = \begin{cases} A * (-\cos(2\pi t/T) + 1) & 0 \leq t < T \\ 0 & \text{otherwise} \end{cases}$$

A constant impact duration T is assumed for any amplitude, which corresponds to an impactor with a linear behavior. Thus, the influence of geometric non-linearities can be assessed without the non-linear behavior of the impactor interfering with this measurement. T was set to 2 ms and the amplitude A was linearly varied over 10 levels from 100 N to 1000 N.

Finite Difference Scheme. Time and space are discretized. We note $u_{l,m}^n$ the transverse displacement at the n^{th} time step and the grid point ($x = h * l; y = h * m$), with h the spacing between two grid points, $k = 1/f_e$ the time step ($f_e = 44100$ Hz).

We used the following conservative scheme, as defined in [26]

$$\delta_{tt}\boldsymbol{u} = -\frac{D}{\rho H}\delta_{\Delta,\Delta}\boldsymbol{u} + \frac{1}{\rho H}l(\mu_{t\cdot}\boldsymbol{\phi}, \boldsymbol{u}) - 2\sigma_0\delta_{t\cdot}\boldsymbol{u} + 2\sigma_1\delta_{t-}\delta_{\Delta}\boldsymbol{u} + \frac{\boldsymbol{J}}{\rho H}F \tag{6a}$$

$$\mu_{t+}\delta_{\Delta,\Delta}\boldsymbol{\phi} = -\frac{EH}{2}l(\boldsymbol{u}, e_{t+}\boldsymbol{u}) \tag{6b}$$

with

$$e_{t+}u_{l,m}^n = u_{l,m}^{n+1}, \qquad e_{t-}u_{l,m}^n = u_{l,m}^{n-1} \tag{7a}$$
$$e_{x+}u_{l,m}^n = u_{l+1,m}^n, \qquad e_{x-}u_{l,m}^n = u_{l-1,m}^n \tag{7b}$$
$$e_{y+}u_{l,m}^n = u_{l,m+1}^n, \qquad e_{y-}u_{l,m}^n = u_{l,m-1}^n \tag{7c}$$
$$\mu_{t+} = (e_{t+} + 1)/2 \tag{7d}$$
$$\delta_{t\cdot} = (e_{t+} - e_{t-})/2k \tag{7e}$$
$$\delta_{t-} = (1 - e_{t-})/k \tag{7f}$$
$$\delta_{tt} = (e_{t+} - 2 + e_{t-})/k^2 \tag{7g}$$
$$\delta_{\Delta} = (-4 + e_{x+} + e_{x-} + e_{y+} + e_{y-})/h^2 \tag{7h}$$
$$\begin{aligned}\delta_{\Delta,\Delta} = (20 &+ 2[e_{x+}e_{y+} + e_{x+}e_{y-} + e_{x-}e_{y+} + e_{x-}e_{y-}] \\ &- 8[e_{x+} + e_{x-} + e_{y+} + e_{y-}] + e_{x+}^2 + e_{x-}^2 + e_{y+}^2 + e_{y-}^2)/h^4\end{aligned} \tag{7i}$$
$$\delta_{xx} = (e_{x+} - 2 + e_{x-})/h^2 \tag{7j}$$
$$\delta_{yy} = (e_{y+} - 2 + e_{y-})/h^2 \tag{7k}$$
$$\mu_{x-,y-} = (e_{x-} + 1)(e_{y-} + 1)/4 \tag{7l}$$
$$\delta_{x+y+} = (e_{x+} - 1)(e_{y+} - 1)/h^2 \tag{7m}$$
$$l(f, g) = \delta_{xx}f\delta_{yy}g + \delta_{yy}f\delta_{xx}g - 2\mu_{x-,y-}(\delta_{x+,y+}f\delta_{x+,y+}g) \tag{7n}$$

J is the interpolation operator. $J_{N_x/2+1,N_y/2+1} = \frac{1}{h^2}$ for the node in the middle of the plate (an even number of elements in length N_x and width N_y is defined), $J_{l,m} = 0$ for the other nodes.

More details are available for the implementation of the scheme in [27].

Loudness Equalization. A total of 30 different sounds were synthesized (3 levels for the thickness of the plate, 10 levels for the strength of the impact). The samples were normalized (their maximum value is set to 1) as follows:

$$s_{norm}^n = \frac{s^n}{\max(|s^n|)} \tag{8}$$

with s^n the original signal and s^n_{norm} the normalized signal at the time step n.

Finally, we proceeded to a loudness equalization. There is no model for the loudness of complex sounds, so it is necessary to probe perception with pairwise comparisons. Comparing all pairs with an adaptive procedure is too time-consuming for a pre-test (435 pairs, with about 1 minute testing time per pair). Furthermore, the task is complex because the sounds to be evaluated have a different temporal evolution. In general, sounds corresponding to a strong non-linear behavior have less energy at low frequencies and are damped more quickly. A simplified procedure consists in choosing the median sample ($H = 2mm$; $A = 500N$) as a reference and compare it with the other sounds. This limits the comparison to 29 pairs and minimizes the difference between the sounds to be compared.

The resulting stimuli are available for online listening [3].

3.2 Participants and Procedure.

Fourteen participants (3 female, 11 male), aged 22 to 50, were tested in this experiment. They all had normal audition and gave consent to participate in the experiment.

Participants were asked to evaluate the strength of the impact and the "suffering" ("souffrance" in French) of the object for each stimulus by moving two sliders on a scale without markers. The 30 sounds were randomly presented through headphones (Sennheiser HD650) in a quiet environment.

3.3 Results

ANOVAs were conducted for the perceived strength of the impact and the perceived "suffering" of the object. Factors were the thickness (1 mm, 2 mm, 3 mm) and the impact strength amplitude (100 N, 200 N, 300 N, 400 N, ... , 1000 N). Results are displayed in fig.1&2.

4 Discussion

There is a direct correlation between the perception of the impact strength and the "suffering" of the object. The evaluation of the strength and the "suffering" are almost identical for all the sounds, although the perceived impact strength is slightly higher in general. This result is also revealed in the feedback from the different participants who said they answered almost the same value for both parameters for most sounds. This reflects the fact that the presence of non-linearities due to large object deformations is used as a perceptual cue to assess the impact strength, which is consistent with the initial hypothesis. This observation is reinforced by the significant differences between the changes in the evolution of the perceived strength regarding the excitation amplitude A for each

[3] https://cloud.prism.cnrs.fr/index.php/s/tx67ywnVvMg21jC

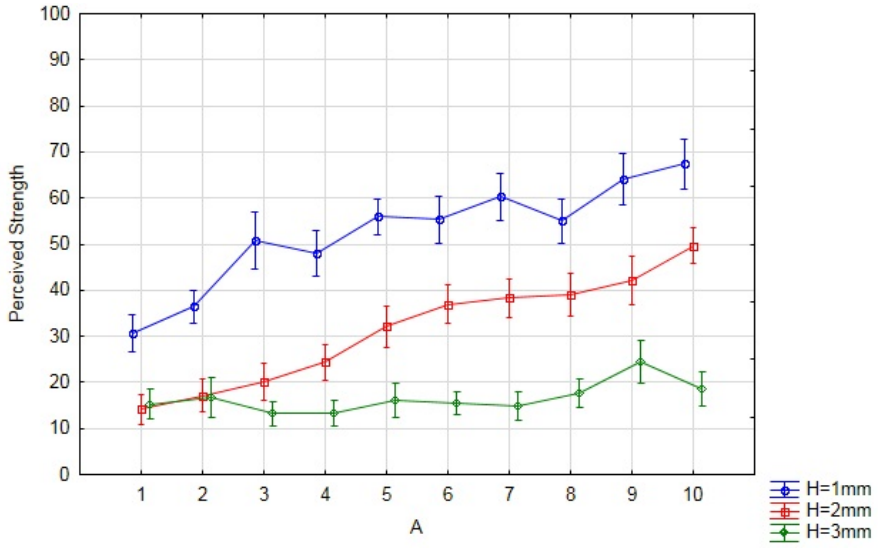


Fig. 1. Representation of responses for the perceived strength of the impact for all samples (percentage of cursor range). The level of the excitation A is presented on the horizontal axis, and each curve represents a level for the plate thickness. Dots denote the average value for all participants, vertical bars denote +/- standard errors.

high H level (the interaction effect AxH is significant $p = 0.00000$). Indeed, the perceived strength for H_3 (almost no non-linear phenomena for the whole range) is relatively constant regarding the level of the impact strength. Conversely, we observe a significant evolution of the perceived strength regarding A for the smallest thicknesses H_1 and H_2, which correspond to sounds that present a progressive appearance of non-linear phenomena when the impact strength increases.

Also, the upper part of the scale remained unused for most participants. Several participants told that they were expecting stronger amplitudes of excitation that would damage or brake the plate. On the other hand, we notice that the strength is never perceived as zero, unlike the "suffering" of the object, and that the value of the perceived strength is always higher than the value of the "suffering" of the object. This is consistent with the idea that the "suffering" of the object only begins when non-linear phenomena begin to appear, unlike the strength that always has to be different from 0 for a sound to occur.

5 Conclusion & Perspectives

The purpose of this study was to investigate perceptual expectations regarding the evolution of the sound radiated by an object with respect to the strength of the impact it undergoes. To this end, we sought to evaluate the effect of non-linear phenomena on the perceived impact strength. This paper focuses on

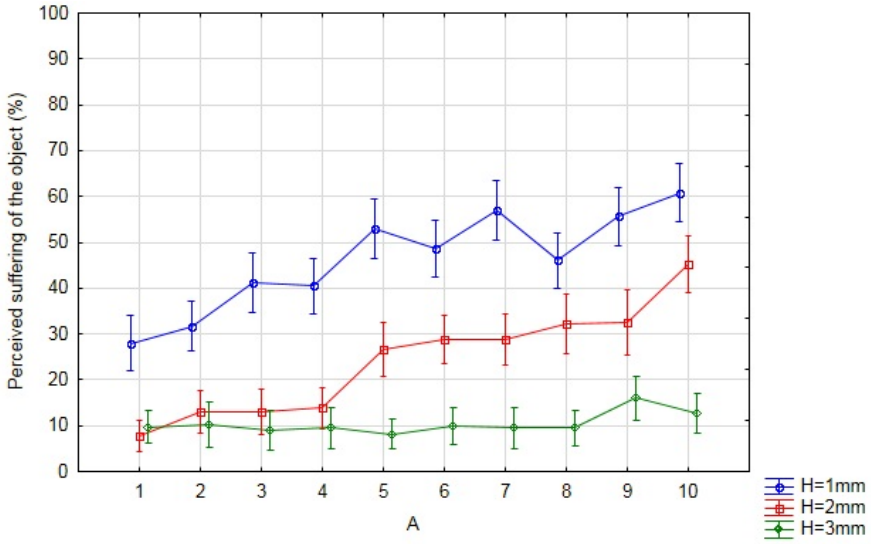


Fig. 2. Representation of responses for the perceived "suffering" of the object for all samples (percentage of cursor range). The level of the excitation amplitude is presented on the horizontal axis, and each curve represents a level for the plate thickness. Dots denote the average value for all participants, vertical bars denote +/- standard errors.

the study of thin plates for moderate vibration amplitudes. A listening test was conducted to evaluate the impact strength and the "suffering" of the object for 30 sounds synthesized by physical modeling of the system, corresponding to 3 plates of different thicknesses impacted with an excitation amplitude ranging from 100 N to 1000 N. The evaluation results show that the perceived impact strength is directly correlated with the occurrence of non-linear phenomena in the case of thin plates.

Further, the deformations are perceived as small for the present sound corpus, since only the lower section of the evaluation scale is used by the subjects. This result encourages us to extend this experiment by modelling the effects of plastic deformations and rupture on the sound radiated by an object.

The next step is to propose signal transformation models corresponding to the different non-linear phenomena, and perceptually relevant controls to improve the expressive potential of percussive sound synthesizers.

References

1. S. Fels, A. Gadd, and A. Mulder, "Mapping transparency through metaphor: towards more expressive musical instruments," *Organised Sound*, vol. 7, no. 2, pp. 109–126, 2002.
2. M. M. Wanderley and P. Depalle, "Gestural control of sound synthesis," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 632–644, 2004.

3. A. Gadd and S. Fels, "Metamuse: metaphors for expressive instruments," in *Proceedings of the 2002 conference on New interfaces for musical expression*. National University of Singapore, 2002, pp. 1–6.
4. R. Hoskinson, K. van den Doel, and S. Fels, "Real-time adaptive control of modal synthesis," in *Proceedings of the 2003 conference on New interfaces for musical expression*. National University of Singapore, 2003, pp. 99–103.
5. S. Conan, E. Thoret, M. Aramaki, O. Derrien, C. Gondre, S. Ystad, and R. Kronland-Martinet, "An intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling," *Computer Music Journal*, vol. 38, no. 4, pp. 24–37, 2014.
6. M. Aramaki, R. Kronland-Martinet, T. Voinier, and S. Ystad, "A percussive sound synthesizer based on physical and perceptual attributes," *Computer Music Journal*, vol. 30, no. 2, pp. 32–41, 2006.
7. M. Aramaki, C. Gondre, R. Kronland-Martinet, T. Voinier, and S. Ystad, "Thinking the sounds: an intuitive control of an impact sound synthesizer." Georgia Institute of Technology, 2009.
8. M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, "Controlling the perceived material in an impact sound synthesizer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 301–314, 2011.
9. R. Kronland-Martinet, S. Ystad, and M. Aramaki, "High-level control of sound synthesis for sonification processes," *AI & society*, vol. 27, no. 2, pp. 245–255, 2012.
10. J. J. Gibson, "The ecological approach to visual perception," 1979.
11. W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.
12. S. McAdams and E. Bigand, "Introduction to auditory cognition," 1993.
13. L. Pruvost, B. Scherrer, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Perception-based interactive sound synthesis of morphing solids' interactions," in *SIGGRAPH Asia 2015 Technical Briefs*. ACM, 2015, p. 17.
14. S. Conan, E. Thoret, C. Gondre, M. Aramaki, R. Kronland-Martinet, and S. Ystad, "An intuitive synthesizer of sustained interaction sounds," in *10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2013, pp. 1045–1050.
15. A. Sackfield and D. Hills, "Some useful results in the classical hertz contact problem," *The Journal of Strain Analysis for Engineering Design*, vol. 18, no. 2, pp. 101–105, 1983.
16. J. Bensa, K. Jensen, and R. Kronland-Martinet, "A hybrid resynthesis model for hammer-string interaction of piano tones," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1021–1035, 2004.
17. A. Stulov, "Hysteretic model of the grand piano hammer felt," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2577–2585, 1995.
18. J. Chowning and D. Bristow, *FM theory and applications: By musicians for musicians*. Hal Leonard Corp, 1987, pp. 150–151.
19. C. Roads and J. Reydellet, *L'audionumérique-3e éd.: Musique et informatique*. Dunod, 2016, pp. 513–514.
20. W. H. Warren and R. R. Verbrugge, "Auditory perception of breaking and bouncing events: a case study in ecological acoustics." *Journal of Experimental Psychology: Human perception and performance*, vol. 10, no. 5, p. 704, 1984.
21. J. O. Smith, "Physical modeling synthesis update," *Computer Music Journal*, vol. 20, no. 2, pp. 44–56, 1996.

22. S. Bilbao, *Numerical sound synthesis: finite difference schemes and simulation in musical acoustics*. John Wiley & Sons, 2009.
23. S. Willemsen, N. Andersson, S. Serafin, and S. Bilbao, "Real-time control of large-scale modular physical models using the sensel morph," in *To appear in Sound and Music Computing*, 2019.
24. D. Rocchesso, R. Bresin, and M. Fernstrom, "Sounding objects," *IEEE MultiMedia*, vol. 10, no. 2, pp. 42–52, 2003.
25. S. Poirot, S. Bilbao, M. Aramaki, and R. Kronland-Martinet, "Sound morphologies due to non-linear interactions: Towards a perceptual control of environmental sound synthesis processes," in *DAFx2018*, 2018.
26. S. Bilbao, "A family of conservative finite difference schemes for the dynamical von karman plate equations," *Numerical Methods for Partial Differential Equations: An International Journal*, vol. 24, no. 1, pp. 193–216, 2008.
27. A. Torin, "Percussion instrument modelling in 3d: Sound synthesis through time domain numerical simulation," 2016.
28. C. Cadoz, A. Luciani, and J. L. Florens, "Cordis-anima: a modeling and simulation system for sound and image synthesis: the general formalism," *Computer music journal*, vol. 17, no. 1, pp. 19–29, 1993.

Musicality Centred Interaction Design to Ubimus: a First Discussion

Leandro Costalonga[1], Evandro Miletto2, Marcelo S. Pimenta3

1 Federal University of Espirito Santo (UFES), Sao Mateus, CEP, 29932-540, Brazil
2 Federal Institute of Rio Grande do Sul (IFRS-POA), Porto Alegre, RS, 90030-041, Brazil
3 Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil
leandro.costalonga@ufes.br

Abstract. All humans share a predisposition for music even those who consider themselves to be "unmusical". Until recently, most scholars were wary of the notion that music cognition could have a biological basis, and this fact reflects on the limited support HCI offers to the design of ubimus technology. In this paper, we present a preliminary discussion on five main aspects of human nature that can be applied to ubimus interaction design including (i) materiality and physicality of musical instruments; (ii) consciousness achieved when skills and challenges are in equilibrium during musical learning; (iii) natural mappings of gestures and movements; (iv) ability to recognize and synchronize with auditory signals; and finally (v) usage of (true) imitation as an strategy to musical learning and communication. It is our intention to point some ideas, concepts and principles that could be used as initial set of interaction design guidelines for improving User eXperience (UX) when developing digital music instruments in ubimus context.

Keywords: musicality, interactive design, new interfaces for musical expression, musical digital instruments, embodied cognition, ubimus.

1 Introduction

Human-Computer Interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers or, as in the context of this paper, digital musical instruments.

Numerous adjectives have been used to describe the different kinds of interfaces that have been developed between computers and humans, including graphical, command, speech, multimodal, invisible, ambient, mobile, intelligent, adaptive, tangible, touchless, and natural [1], several of which have been tried for musical tasks [2] and, with a few exceptions, unsuccessfully. To be successful the designer's focus needs to go beyond the item in development to include the way users will interact with it considering not only the users' needs but also user's limitations and its contexts of use. This approach to the design of interactive products and services is known as Interaction Design (IxD).

A relevant change of paradigm in HCI was Weiser's [3] vision of ubiquitous technology (UbiComp) [1]. The readings of UbiComp concepts and technology by fields such as Music and Arts led to the birth of subfield so called Ubiquitous Music (UbiMus, concerned with ubiquitous systems of human agents and material resources that afford musical activities through creativity support tools [4]. In practice, ubimus is music (or musical activities) supported by ubiquitous computing concepts and technology.

The growing interest from the Computer Music community in digital and computer interfaces used for music performance is a clear indication that Computer Music (CM) researchers have become aware of the importance of HCI studies. Indeed, interfaces for musical performance present immense challenges for HCI: since such interfaces provide the interaction between a performer and a computing system (involving several complex cognitive and motor skills), they make computers turn back toward being things that fit our physical, embodied natures, rather than only operating in the realm of symbolic processing. Therefore, such interfaces for musical performance should be designed around human inputs and outputs rather than computer ones [5] and this claim remains true in the context of ubimus.

One of the most pertinent questions for the 21st century is how these increasingly intelligent and invasive technologies will affect our minds. The term "extended mind", has been used in order refer to notion that our cognition goes beyond our brains and suggests that individuals may allow their smart devices to do their thinking for them [6]. The ability to rely on the external mind might have detrimental consequences to cognition [7] because humans are ''cognitive misers'', meaning that people tend to eschew costly analytic thought in favor of comparatively effortless intuitive processing. There is evidence suggesting that users that make high use of smartphones are genuinely lower in analytical reasoning leaning more towards an intuitive decision-making[6]. Based on that premise, is it possible to argue that the use of ubimus technology could also shape the way we think music or has any detrimental effect on our musicality? Despite some evidences, it would be premature state that. However, it is possible to approach this issue from a different angle, that is, researching about the cognitive and biological traits involved in musical thinking and applying them to the design of new digital instruments.

The goal of this paper is to present a preliminary discussion on five main aspects of human nature that can be applied to ubimus interaction design. After introducing some concepts related to musicality (section 2), the section 3 discusses some guidelines related to how to integrate such musicality concepts to HCI, in particular to Interaction Design of Digital Musical Instruments (DMI).

Note that, in the context of the work, we will refer to "music" as a product of musicality, a social and cultural construct of humankind based on the presence of several cross-cultural similarities. "Musicality", in turn, can be defined as a natural, spontaneously developing set of traits based on and constrained by our cognitive and biological system. In summary, music in all its variety can be defined as a social and cultural construct based on that very musicality [8], as will be discussed in the next section.

2 Musicality: Cognitive and Biological Musical Traits

Musicality is our universal birth right as human beings. Everyone is a skilled musical listener, even those who consider themselves to be "unmusical" have an implicit knowledge of the musical forms and styles of their culture (even if they cannot be expressed explicitly) [9]. All humans share a predisposition for music, just like we have for language. To recognize a melody and perceive the beat of music is an example of a trait based on and constrained by our cognitive abilities and their underlying biology. Even infants are sensitive to such features, which are common across cultures [10]. Indeed, we find music in all cultures and time periods.

Until recently, most scholars were wary of the notion that music cognition could have a biological basis. Music was viewed as a cultural product with no evolutionary history and no biological constraints on its manifestation typically. This explanation is based on the belief that music has not been around long enough to have shaped perceptual mechanisms over thousands of generations. Moreover, in contrast to speech, musical knowledge is acquired relatively slowly and not equally by each individual [11]. More recently, studies has indicated that our capacity for music has an intimate relationship with our cognition and underlying biology, which is clear when the focus is on perception rather than production. [9].

Comparative research shows that although music itself may be specifically human, some of the fundamental mechanisms that underlie human musicality are shared with other species. Research has revealed detailed commonalities of bird or whale "song" with our own music, as well as some specific homologs such as the drumming behavior regularly engaged in by our nearest relatives, the chimpanzees [12].

For Darwin, music had no survival benefits but offered a means of impressing potential partners, thereby contributing to reproductive success, although there are divergent studies on that matter (see other reported purposes for music in [7]).

All in all, consensus is growing that musicality has deep biological foundations, based on accumulating evidence for the involvement of genetic variation [13]. Recent advances in molecular technologies provide an effective way of exploring these biological foundations, such as the association studies of genome aiming to capture the polymorphic content of a large phenotype population sample.

We are convinced that a well-applied knowledge of some aspects of human biological nature can be beneficial to the ubimus interaction design, in particular to the design of more natural digital music instruments. By adopting such biological foundation we have the chance to rethink the limited support HCI offers to the design of ubimus technology. In that sense, some guidelines are discussed in the next section.

3 Guidelines on Applied Musicality to HCI

Historically, the HCI perspective is often related to concepts, models, methods and techniques in which the final intention is to optimize the user interface (UI) around how people can, want, or need to work, rather than forcing the users to change how they work to accommodate the system or function. In fact, the HCI perspective is based on the needs of the users and, for that, we need to know them, their goals and tasks. In

other words, to adopt the HCI perspective in computer music converges towards the central idea that, to design more useful and usable systems, we must better understand the users and the tasks they undertake as well as apply our understanding of these tasks in the design process. Such approach (known as User-Centered Design, UCD) is also valuable for the design of digital devices-applications-gadgets for music.

People usually had an "emotional" affection towards their acoustic instruments, and they bonded with its character. It does not mean they are perfect! Normally, the learning of an acoustic instrument starts at a very young age when things are more likely to be taken for granted - if one cannot play the instrument properly, it is felt as not an imperfection of the instrument design itself but one's own fault. The fact that acoustic instruments seem to have existed forever makes people less likely to step back and actively criticize their instrument of choice [14]. This is very different in regard to people's feelings about their digital instruments and, a big part of it, is due to the user experience (UX).

As the name suggests, UX design is about designing the ideal experience of using a service or product. It is about the way people feel about a product and their pleasure and satisfaction when using it, looking at it, holding it, etc. To achieve this, the users must be in the center of the designing process (UCD approach): they must not only be listened but also be involved. Overall, it is essential to take into account what people are good and bad at, both in a motor and cognitive level. For that reason, HCI has been always interconnected with the fields or ergonomics and cognitive sciences [1]. In the next sections it is presented five aspects of human nature that could be applied to ubimus interaction design, described as follows.

3.1 Materiality and Physicality of Musical Instruments

The production of structured communicative sounds by striking objects with limbs, other body parts, or other objects appears to constitute a core component of human musicality with clear animal analogues [12]. Musicians lost the very subjective feeling of actually objectifying the sound by laying it down as incomprehensible 0' s and 1' s distributed on a spinning metal plate. All the buttons, wiring, switches, knobs and faders for controlling the recording and mixing of audio— when remediated to a computer interface— had to be accessed through either a 3-button mouse or the standard QWERTY-keyboard [2]. Although these devices are tangible, in the same way as the early MIDI-controllers were, they lack clear physical relation between the physical and digital representation.

Acoustic instruments contain properties due to their materiality that limits its sound scope. The interfaces of acoustic instruments are relatively simple to understand. There are "natural" mappings between the exertion of bodily energy and the resulting sound, either by plucking a string or hitting a drum with a stick [14]. Is not only about the audible feedback, it is also about the haptic feedback determining by the very shape and materiality of acoustic instruments that is missing in the digital counterparts [2]. In all acoustical instruments, the human voice included, the depth and speed of a vibrato is proportional by the amount of force applied to the instrument. A vibrato on a guitar, for instance, differs dramatically from digital musical instruments.

Playing digital instruments seems to be less of an embodied practice (where motor-memory has been established) as the mapping between gesture and sound can be changed so easily. In a survey with over 200 musicians, Magnusson & Mendieta [14] reported that participants expressed the wish for more limited expressive software instruments, i.e. not a software that tries to do it all but "does one thing well and not one hundred things badly". Software instruments, including sonic programming languages, presents endless possibilities and are too broad to get to know it thoroughly.

Another relevant aspect regarded the materiality of musical instruments is concern with its perenniality and durability. It is frustrating having to deal with updates, fixing, compatibilities, and the overall uncertainty of the continuation of commercial digital instruments or software environments. Acoustic instrument does not have a due date and it will not become updated with the next year's release; such credibility is an important value of the UX.

Acoustic instruments also have its own interaction issues. When the sound production mechanics are tied with the interface, the instruments may not prioritize the ergonomics which leads to discomfort, errors, and injuries. Modern technology has made possible to separate the user interface from the sound production mechanism of an instrument. As a result, the interface could be optimized for usability/UX and the sound production of the instrument could be separately optimized, without the need for constraints such as keeping the pipes close enough together for a musician to be able to reach each pipe or each finger hole with his or her fingers [15]. Despite the advantages of the separation of user interface from sound-producing medium, a price payed for this separation is the loss of physicality. The synthesizer fundamentally changed the haptic aspects of musical performance, by essentially eliminating it. At the same time, however, the synthesizer also augmented the sonic vocabulary, paving the way for new musical expression through new sounds and new timbres [2].

Musical instrument must sound good, but also must be not too difficult to play, relatively durable, and more or less affordable [16]. Manipulating an acoustic musical instrument for an extended period of time will almost certainly lead to injuries resultant of the accumulation of micro traumas. Actually, 48-66% of string players report injuries serious enough to interfere with their ability to perform [17].

We summarize now some guidelines related to materiality and physicality of musical instruments: a) to invest in a durable and sustainable design using good quality materials that evokes desire in the users – not a disposable toyish design; In fact, a redesigned augmented instrument would be preferred; b) to limit the features of the instrument in a way that it would have an easy learning curve but incorporates deep potential for further explorations, so it will not to become a boring; c) to avoid the temptation to just copy the interface of the acoustic instrument on a new digital counterpart. It is positive to incorporate familiar materials (i.e. guitar strings) and make use of the known haptic vocabulary (i.e. the vibration of the string when plucked). Do it after ergonomic studies and evaluation with actual users to avoid possible injuries.

3.2 Appease Skill, Challenge, and Assistance

In a traditional musical context, it is required decades of regular practice, estimated at 10,000 hours, to become a skilled musician [18]. It is just too difficult and, yet, it

seems that learning a musical instrument is still a very desirable thing! Acoustic instruments have longer lifetime, which makes practicing them more likely a continuous path to mastery [14]. A challenging instrument to master will definitely set apart the skilled ones and reward them with the rightful status and admiration, which can very motivational. As a drawback, acoustic instruments might not be very accessible requiring specific physical attributes from the players.

Musical learning is done also for pleasure. Csikszentmihalyi [19] argued that an intrinsically rewarding state of consciousness (negentropic state) is achieved when skills and challenges are in equilibrium. As a beginner, it can be both exciting and daunting to observe a virtuoso playing. Ideally, the interaction with a digital musical instrument should provide support for different users' profiles and skill levels allowing a gradual and consistent growth of the musical abilities, both motor and cognitive.

The translation of musical thoughts into actions usually requires, not only talent, but also a fair amount of practice. These actions can be either motor-related, such as playing a fast passage on a traditional musical instrument, or cognitive-related, such as structuring and developing a musical algorithm. To amend that, several commercial musical products offer musical libraries composed by samples, patches, setups, rhythmic patterns, riffs, etc. that facilitates the process creating music. These building blocks encapsulates a set of complex tasks that can be looked into detail whenever the user feels prepared to do so. This is a good thing, but it can be also very limiting expression-wise. In addition, picking and choosing the right "ingredients" for musical receipt presumes that you are familiar with the all the ingredients and how they will blend. This is a problem with every larger musical library. How to get the right content without wasting time?

Barr et al. [6] studies suggest that people who think more intuitively when given reasoning problems were more likely to rely on their connected devices, suggesting that people may be prone to look up information that they actually know or could easily learn but are unwilling to invest the cognitive cost associated with encoding and retrieval. Music is still believed to be mostly a matter of the "intuitive" right brain – the avatar of emotion and creativity [11]. If that so, chances are the user will never look into the building blocks since one might choose not to engage in costly elaborative encoding, as they know that knowledge can be procured externally. Therefore, building blocks strategy might not be ideal for some tools for musical learning. Another interpretation is that, if people relies on their digital knowledge base to take action or even pass over the control of decision making to the algorithms, then the musical interface itself might be seen as secondary for this particular group profile.

We summarize now some guidelines related to conciliate skill, challenges and assistance: a) improved search engines based on musicality: vocalization/humming searching and/or rhythmic searching based on physical relation between the physical and digital representation (i.e. drumming). Also, the results could be sorted based on the user's skill level. Mainly, do not offer too much too soon nor too little too late; b) Intelligent recommendation systems: autocomplete musical phrases, intelligent harmonizers and timbre recommendation system based on a particular musical style; c) performance-related assistance such as adaptive accompaniment systems or an intelligent "autopilot" mode. Whatever the level of the assistance offered, the design must also reflect a well-

thought balance between assistance and intrusiveness, since there is evidence that more knowledgeable individuals are less likely to enjoy it and might want to switch it off.

3.3 Gestures and Movements

It has been already mentioned the human propensity to generate percussive sounds using limb movements ("drumming") as a core component of human musicality [12]. Interfaces that use our full body movements, for example gestural interfaces, are supposedly more "natural" than mouse and keyboard interfaces because body movements come naturally to us. The reality is that bodily interfaces can suffer from problems associated with traditional interfaces (i.e. the difficulty of remembering gesture) as well as new problems such as the ephemerality of gestures and lack of visual feedback. Natural User Interfaces (NUI) are only natural if they use our existing sensorimotor body movement skills to enable us to interact with the physical world. They must be thought as tapping sensorimotor skills, not representation manipulation skills [20]. Therefore, findings from cognitive science are vital to understand how these embodied skills work.

Todd [21] defends the principle that performance, perception of tempo and musical dynamics are based on an internal sense of motion. This principle reflects upon the notion that music performance and perception have their origins in the kinematic and dynamic characteristics of typical motor actions. In order to sound natural in performance, expressive timing must conform to the principle of human movement [22]. Baily [23] even argues that the performer's internal representation of music is in terms of movement, rather than sound. A shared assumption is that we experience and make sense of musical phenomena by metaphorically mapping the concepts derived from our bodily experience of the physical world into music, that is, a tight mapping between our movements and changes to our sensory input. For instance, changes in sound energy makes sense when mirrored in the subject's action-oriented ontology.

The process forms the basis of an appreciation of music which is strongly based on body movement, and to which cerebral appreciation and interpretation can be added [24]. Bevilacqua et al. [25] proposed a system that allows users to easily define their own gestures and they do so by acting out those gestures while listening to the music to be controlled. This means that gestures are not limited to a set of pre-defined symbolic gestures, but it can be defined based on what feels natural to a particular user. In addition, gestures did not trigger an action when recognized by the system but instead it continuously controls the production of sound throughout the time.

Corporeal articulation is also related to musical expressiveness and can be seen as indicators of intentionality (studied in terms of this mirroring process) [24]. In general terms, movement in response to music is often seen as a gestural expression of a particular emotion (sadness, happiness, love, anger) that is assumed to be imitated by the music [26]. Note, however, that the expression of emotion is only one aspect of corporeal involvement with music since corporeal articulations can also be used to annotate structural features, such as melody, tonality, and percussion events.

Another aspect to be considered regarded to the interrelationship between sound and movement is related to the ideomotor principle, which is the tendency that people have

to move in synchrony with auditory rhythm. The effect is clearly observable in the tendency to tap along with the beat of the music [28]. In this regard, Knuf et. al. [27] ran a comprehensive study on ideomotor actions and verified that movements did not always occur without awareness, but they did occur without awareness of voluntary control. They have also provided clear evidence that people do tend to perform the movements they would like to see (intentional induction) whereas results are less clear with respect to the perceptual induction (movements that people actually see). Perceptual induction could only be verified thru non-instrumental effectors, in their experiment, the effect appeared for both head and foot. For hand movements (in that case, the instrument effectors), intentional induction is much more pronounced than perceptual. Note, however, that these experiments did not made use of auditory signals.

We summarize now some guidelines related to natural mappings of gestures and movements: a) Movements are good for controlling rhythmic information, both during the performance and for data input. Hand and upper body limbs (instrumental effectors) movements are preferred; b) Movement are also good to control dynamic and other parameters of musical expression. Other forms of body language, such as face expression detection, may also leads to good results; c) Do not impose a set to gestures and movements that does not conform to our existing sensorimotor body movement skills. Movements that represents manipulation skills should be avoided (i.e. air guitar). It might be a good idea to observe and learn personal movements of a particular user during a musical performance; d) Perceptual induction could be used as a way to get engagement feed-back from the audience.

3.4 Entrainment, Synchronization, and Beat Perception

Honing [8] defends that regularity/beat perception is another of the human's innate musical traits. Beat perception and synchronization refers to the capacity to extract an isochronic beat and synchronize to it [29]. Cate et al. [30] goes beyond studding in different species both the ability to recognize the regularity in the auditory stimulus and the ability to adjust the own motor output to the perceived pattern. Although rare, some other species do share some similar entrainment capabilities as humans, nevertheless, for us, this capacity has virtually no use outside music, dance, and drill [12].

Human rhythmic abilities obviously did not arise to allow people to synchronize to metronomes but rather to the actions of other humans in groups, known as social synchronization. Thus, the concept of mutual entrainment among two or more individuals should be the ability of central interest rather than BPS to a mechanical timekeeper [12]. In nature (i.e., outside the laboratory), the mutual entrainment of two or more individuals by such a mechanism obviously cannot occur unless they themselves are capable of producing the evenly paced entraining stimulus. That is, they must be capable of endogenously producing isochronous behavioral output without prior input, strictly speaking, a propensity to spontaneously produce isochronous behavioral cyclicity of some kind (such as clapping, stomping, or drumming) within the tempo range of its predictive timing mechanism [31].

The most sophisticated form of synchronization involves beat-based predictive timing, where an internal beat is tuned to the frequency and phase of an isochronous time-

giver, allowing perfect 0-degree phase alignment. This stimulus makes the very next beat in the sequence predictable, allowing the timing mechanism to align—or latch—its produced behavioral to the stimulus with zero, or even small negative (anticipatory), phase lag, typical of human sensorimotor synchrony [32]. Because of reaction time limitations, it cannot therefore be based on responding to that stimulus event. Instead, it requires a predictive (anticipatory) and cyclical motor timing mechanism that takes an evenly paced stimulus sequence as input. Naturally, reaction times to predictable stimuli are shorter than those to unpredictable ones, hence preparatory cues such as "ready, steady" or a drummer's count down allow quicker responses.

Normally, humans have a preferred tempo centered on 2 Hz (500 ms or 120 BPM) where precision is highest (standard deviation of a 2..5%) and around which the tempo of rhythmic music concentrates. The precision substantially declines above 900ms (67 BPM) interval length [31], requiring a cognitive strategy of mentally subdividing long periods (so-called subitizing) to maintain precision [33].

We summarize now some guidelines related to ability to recognize and synchronize with auditory signals: a) Make use of auditory predictive cue to synchronized-demanding actions. Sound has been largely used as a way to call for attention in interactive systems, however in interactive systems for music performance this must be used very carefully to not disrupt the main purpose of the system. That said, it is possible to lean on our innate ability encode metrical information, such as rhythm, to call for the user attention when some action is required. For example, we can use syncope or other rhythmic cues as a feedback for the user. In part, this is a very common strategy used by DJ's when mixing electronic musical styles. b) Offers support to rhythm-related activities when BPM drops below 70 (i.e. a visual representation of the tempo). c) People are good at synchronizing with other people, not machines. Tangible User Interfaces are a good option since they tend to provide support for more than one person to explore the interface together at the same geographical location, therefore offering the chance for them to make the most of their innate ability to entrain.

3.5 Imitative Behavior for Musical Communication and Learning

(True) imitation is innate. It is well-developed in humans being observed in newborns babies both for fostering learning and for yielding pleasure. There is a distinction between imitation that copies the task structure and hierarchical organization, and imitation that copies movements. True imitation focuses on the goal, that is, the execution of the action as a function of the goal [34]. Learning to play a musical instrument is, therefore, a typical example of true imitation. It draws on the ability of the student to focus on what is essential in the teacher's example. Even if the instrument is not the same it is still possible to imitate particular behaviors and playing styles because the student has more of a focus on the goals and less of a focus on the precise movements.

The role of mirroring in music education has been confirmed by a brain imaging study that revealed that mirror neurons encoded decomposed elementary motor components [35]. When the action to be imitated corresponded to an elementary action already present in the mirror neuron system the act was forwarded to other structures and

replicated. In that case, no learning was needed. Other forms of imitation observed in a musical context are: imitation skills, musical figures, imitation of symbols, imitation of moving sonic forms (corporeal imitation), and imitation of group behavior (allelo-imitation). Along this line, perceptual induction (ideomotor) can also be considered a special case of imitation and implies that people tend to perform the movements they see. That is, the spectator's actions would tend to repeat what is seen in the scenario.

We summarize now some guidelines related to usage of (true) imitation as a strategy to musical learning and communication: a) Implementation of comparison mechanisms that allows casual players to match their performances against more skilled ones; b) Tutorials with professional players (setting an idol reference) in immersive learning environments (multi-angle) may contribute to a faster development of the required motor skills; c) adoption of elementary motor gestures and movements in an effort to reduce motor-related learning making sure, however, the movements are meaningful and not conflicting with one another or stablished patterns.

4 Final remarks

Over the centuries, the way music has been done, shared, and learned has remain unaltered. Students observe their teachers' movements performed on musical instruments, the intentionality of their actions, the communication strategies involved in this social activity, and all the other facets that guarantees the transmission of this knowledge to the next generation. Humans evolved for that and excel at leaning and playing music this way. However, with the advance of technology and HCI, new forms doing music has emerged supported by technologies such as Computer-Supported Cooperative Work (CSCW), ubimus, and Internet of Musical Thing (IoMusT) [37]. These are very exciting lines of researching, but the big question is: are we cognitively equipped to make the most of it? Could we thrive in this new way of making music or it will become "yet-another" short-lived cool interface for music making? Is this new technology really paying attention to the way we do things, taking into consideration what we are good and bad at? Clearly, there is an intrinsic exploratory value in these initiatives, and it will certainly lead to unpredictable artistic outcomes, but this is insufficient to answer the questions above.

This paper discussed the relevance of applying innate human abilities (musicality) in the design of new digital musical interfaces. Five aspects were presented and for every aspect discussed some preliminary guidelines to the design of new musical instruments were provided. This paper does not claim to have covered all the innate abilities used for musical activities or to have presented the best possible application for them. It is just an attempt to demonstrate that some the basis of HCI regarded to cognitive science and ergonomics might have not been considered in the current designs of digital musical instruments and ubimus technology as a whole. It is yet to be demonstrated the effectivity of each of these reported aspects, as well as others, through usability and UX concerns. We hope our preliminary set of guidelines could help designers and researchers to improve UX related to their musical systems, applications and gadgets or – at least – to increase the discussion about this important subject.

References

1. Preece, J., Rogers, Y., Sharp, H.: Interaction design : beyond human-computer interaction. Wiley, Chichester, UK. (2015)
2. Bech-Hansen, M.: Musical Instrument Interfaces. A Peer-Reviewed J. About. 1, 10 (2013)
3. Weiser, M.: The Computer for the 21 st Century. Sci. Am. 265, 94–105 (1991)
4. Keller, D., Lazzarini, V., Pimenta, M.S. eds: Ubiquitous Music. Springer International Publishing, Cham (2014)
5. Miranda, E.R., Wanderley, M.M., Kirk, R.: New digital musical instruments : control and interaction beyond the keyboard.
6. Barr, N., Pennycook, G., Stolz, J.A., Fugelsang, J.A.: The brain in your pocket: Evidence that Smartphones are used to supplant thinking. Comput. Human Behav. 48, 473–480 (2015)
7. Carr, N.: The Shallows: what the internet is doing to our brains. W.W. Norton & Company, New York, New York, USA (2010)
8. Honing, H., Ploeger, A.: Cognition and the Evolution of Music: Pitfalls and Prospects. Top. Cogn. Sci. 4, 513–524 (2012)
9. Honing, H.: The origins of musicality. The MIT Press, Cambridge, Massachusetts, USA (2018)
10. Trehub, S.E., Becker, J., Morley, I.: Cross-cultural perspectives on music and musicality. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 370, 20140096 (2015)
11. Repp, B.H.: Some cognitive and perceptual aspects of speech and music. In: Music, Language, Speech and Brain. pp. 257–268. Macmillan Education UK, London (1991)
12. Fitch, W.T.: Four principles of bio-musicology. Philos. Trans. R. Soc. B Biol. Sci. 370, 20140091–20140091 (2015)
13. Oikkonen, J., Onkamo, P., Järvelä, I., Kanduri, C.: Convergent evidence for the molecular basis of musical traits. Sci. Rep. 6, 39707 (2016)
14. Magnusson, T., Mendieta, E.H.: The acoustic, the digital and the body. In: Proceedings of the 7th international conference on New interfaces for musical expression - NIME '07. p. 94. ACM Press, New York, New York, USA (2007)
15. Mann, S., Steve: Natural interfaces for musical expression. In: Proceedings of the 7th international conference on New interfaces for musical expression - NIME '07. p. 118. ACM Press, New York, New York, USA (2007)
16. Manchester, R.A.: Musical instrument ergonomics. Med. Probl. Perform. Art. 21, 157–159 (2006)
17. Shan, G.B., Visentin, P.: A quantitative three-dimensional analysis of arm kinematics in violin performance. Med. Probl. Perform. Art. 18, 3–10 (2003)
18. Ericsson, K.A.: The Role of Deliberate Practice in the Acquisition of Expert Performance. Psychol. Rev. 100, 363–406 (1993)
19. Csikszentmihalyi, M.: Imagining the self: An evolutionary excursion. Poetics. 21, 153–167 (1992)
20. Gillies, M., Kleinsmith, A.: Non-representational interaction design. In: J., B. and A., M. (eds.) Contemporary Sensorimotor Theory. Studies in Applied Philosophy,

Epistemology and Rational Ethics. pp. 201–208. Springer, Cham (2014)
21. Todd, N.P.M.: The dynamics of dynamics - a model of musical expression. J. Acoust. Soc. Am. 91, 3540–3550 (1992)
22. Honing, H.: The final ritard: On music, motion, and kinematic models. Comput. Music J. 27, 66–72 (2003)
23. Baily, J.: Music structure and human movement. Music. Struct. Cogn. 237–58 (1985)
24. Leman, M.: Embodied music cognition and mediation technology. MIT Press, Cambridge, Massachusetts (2008)
25. Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., Rasamimanana, N.: Continuous realtime gesture following and recognition. In: S., K. and I., W. (eds.) Lecture Notes in Computer Science. pp. 73–84. Springer, Berlin, Heidelberg (2009)
26. Friberg, A., Sundberg, J.: Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. J. Acoust. Soc. Am. 105, 1469–1484 (1999)
27. Knuf, L., Aschersleben, G., Prinz, W.: An analysis of ideomotor action. J. Exp. Psychol. Gen. 130, 779–798 (2001)
28. Hallam, S., Cross, I., Thaut, M.: The Oxford handbook of music psychology. Oxford University Press, Oxford, UK (2005)
29. Patel, A.D.: Musical Rhythm, Linguistic Rhythm, and Human Evolution. Music Percept. 24, 99–104 (2006)
30. ten Cate, C., Spierings, M., Hubert, J., Honing, H.: Can Birds Perceive Rhythmic Patterns? A Review and Experiments on a Songbird and a Parrot Species. Front. Psychol. 7, 730 (2016)
31. Merker, B., Morley, I., Zuidema, W.: Five fundamental constraints on theories of the origins of music. Philos. Trans. R. Soc. B Biol. Sci. 370, 20140095–20140095 (2015)
32. Fraisse, P.: Rhythm and tempo. Psychol. Music. 2, 149–180 (1982)
33. Repp, B.H.: Self-Generated Interval Subdivision Reduces Variability of Synchronization with a Very Slow Metronome. Music Percept. 27, 389–397 (2010)
34. Byrne, R.W., Russon, A.E.: Learning by imitation: a hierarchical approach. Behav. Brain Sci. 21, 667–84; discussion 684-721 (1998)
35. Buccino, G., Binkofski, F., Riggio, L.: The mirror neuron system and action recognition. Brain Lang. 89, 370–376 (2004)
36. Keller, D., Lazzarini, V., Pimenta, M.S.: Ubimus Through the Lens of Creativity Theories. In: Keller, D., Lazzarini, V., and Pimenta, M. (eds.) Ubiquitous Music. pp. 3–23. Springer, Cham, Cham, UK (2014)
37. Turchet, L., Fischione, C., Essl, G., Keller, D., Barthet, M.: Internet of Musical Things: Vision and Challenges. IEEE Access. 6, 61994–62017 (2018)

A Soundtrack for *Atravessamentos*: Expanding ecologically grounded methods for ubiquitous music collaborations

Luzilei Aliel[1 4 [0000-0001-6728-6451]], Damián Keller[2 4 [0000-0002-0866-3066]], Valeska Alvim[3]

[1] University of São Paulo
[2] NAP (Federal University of Acre and Federal Institute of Acre)
[3] Nois da Casa (Federal University of Acre)
[4] Ubiquitous Music Group
luzaliel@usp.br; dkeller@ccrma.stanford.edu

Abstract. We tackle the creative development of a soundtrack for a video-dance artwork entitled *Atravessamentos*. Our exposition follows two strands: 1) The processes of creative sharing - involving the production of a collaborative artistic work; 2) The development of a multimodal artistic proposal through the application of an ecologically grounded approach. An overview of ecologically grounded creative practices serves to set the context for a discussion of the artistic strategies employed in this project. The contributions include: an expanded notion of *shifting*, incorporating the concept of *cognitively dissonant strategies* and the introduction of *acoustic-instrumental synthesis* as an ecologically grounded technique. We discuss the implications of the application of these strategies within the context of current creative practices in ubiquitous music.

Keywords: ubiquitous music; ecological creative practice; cognitive dissonance.

1 Introduction

Ecologically grounded creative practices target creative actions as by-products of action-perception cycles (20). The ecological perspective rests on the empirical evidence provided by embedded-embodied cognition (13; 14), incorporating recent technological advances such as the application of algorithmic techniques for music making (5) and methods developed within the context of ubiquitous music research (22). Rather than limiting itself to the manipulations of symbols, music creation is thought as interaction among agents and objects situated in specific contexts (16), involving the exploration of local resources and fostering the adoption of open-ended creative forms to encourage musicians (26) and audience (19) to fully engage as creative partners. The environmental resources are treated as central components of eco-based creative practices, highlighting the importance of the *place* factor (27). Current ecological approaches rely on (1) intense social interactions among musicians and non-musicians, (2) usage of everyday settings, and (3) open-ended and

exploratory strategies, thus unveiling the need for alternative methods to support aesthetic decision making.

Focused on the relationships between the creative potential of the local settings and the decision-making strategies, within the ecologically grounded perspectives there are at least two frameworks to deal with sonic resources: fixed and adaptive sonic ecologies (17). Fixed sonic ecologies enforce a strict separation between the decision-making processes and the ecological niche. Thus, decision making does not depend on the local material resources or on the cognitive scaffolding provided by the environmental cues. Complementarily, when the creative products are shaped through processes tied to the local resources, the sonic ecologies are labelled adaptive. This type of ecology hints at a process of mutual adaptation: the ecological niche determines the material resources used for decision making and simultaneously the process of decision making impacts the ecological niche. In eco-oriented terminology, an adaptive sonic ecology can be defined as *a habitat where agents and objects interact producing creative sonic by-products that depend on the local materials and on the behavioural resources* (17). The behavioural processes that occur in both fixed and adaptive ecologies feature four operations: *constrain*, *expand*, *shift* and *nil.* Constrain involves the reduction of resources. Expand targets the exploratory actions needed for resource foraging. Shift combines the two former operations yielding behavioural adjustments to the ensuing conditions. For behaviours that have no material impact on the resources, Jones and coauthors (15) propose the *no operation* label.

In an effort to integrate the research agendas of the ubimus perspectives and the ecologically grounded creative practices, Keller and Lazzarini (20) propose ecologically oriented ways to conceptualize ubimus research. A case in point is their attempt at defining ubiquitous music as a process involving four components, 1) human agents and 2) material resources that 3) afford musical activities enabled by 4) creative support ecosystems.

- Component 1 of the definition refers to the factors that shape hominid evolution. Two of the currently most influential perspectives on evolution theory are the social brain hypothesis (8) and the niche construction theory (28). The former highlights the importance of social interaction mechanisms for survival. The ability to perceive and to predict the intentions of others may have been a key demand of early hominid interactions, requiring a high investment of cognitive resources[1]. The niche-construction perspective supports the embedded-embodied approach to cognition, pointing to a mutual relationship between the adaptive behaviours and the local habitats. The impact of the biological and cognitive adaptations to the environment on the creative behaviours is an open question. Much theoretical and empirical ubimus work remains to be done.
- Component 2 of the ubimus definition – material resources – is highlighted by the creative projects that make use of environmental features to constrain

[1] In the philosophical and cognitive-science literature, this phenomenon is also known as the *Theory of Mind*.

the behaviours of the agents or to generate and organise other material resources. Ubimus systems that make heavy use of environmental features can be classified as multimodal ecologies. For example, ecological modelling proposes the implementation of decision helpers based on the objects' sonic behaviours within terrestrial settings (24). Another possible strand of research features the use of material resources as knowledge sharing mechanisms. This perspective is supported by the recent advances in DIY and DIT[2] techniques (3; 25), targeting the development of prototyping strategies accessible to participants with intermediate musical knowledge. This line of research may also be boosted by the emergence of the Internet of Musical Things (IoMust - 33), encompassing the use of low-cost everyday devices as resources for creative music making.

- Both component 1 and component 2 interact to shape component 3 – *affordances.* The concept of affordance – that is, opportunities for action provided by the features of the environment as perceived by the active agent – was coined by psychologist Gibson (14). Nevertheless, there are multiple examples of problematic usage in the literature on human–computer interaction (see 18 for a critical discussion of the implications and limitations of the concept of affordance for technologically based creative practice). Affordances remain problematic until today. For this reason, Keller et al. (22) have suggested the alternative label *relational properties* to describe the design qualities that emerge from interaction (23). In short, material relational properties arise from interactions with physical objects. Social relational properties emerge from interactions among stakeholders. And formal relational properties[3] include cognitive simulations and conceptual operations handled through offline cognitive resources.
- The fourth component of the ubimus definition expands the notion of the creative tool to tackle systemic relationships. Keller and Lazzarini (21) point out that adopting the tools as a central goal of a research agenda on creativity may be problematic. Artistic activities have been described as self-reflective (7). In other words, the objectives of the activity are usually established during the act rather than a priori. Ubimus ecosystems function as technological hubs that demand the integration of audio and interaction tools. Some of these ecosystems may be reconfigured according to the needs of the stakeholders through rapid prototyping techniques. Thus, ubimus systems may enhance the users' creative potential by providing access to previously unavailable material or social resources. A challenge faced by ubimus designers is to provide intuitive tools for diverse creative tasks without reducing the sustainability of the creative processes.

[2] DIY: *do it yourself*. DIT: *do it together*

[3] Keller and coauthors (2014b) suggest that formal relational properties are decoupled from the processes that take place during synchronous interactions. This proposal still needs full-fledged empirical studies.

Multimodality in ubimus is closely related to components 2, 3 and 4. Material resources, musical activities and creative support tools foster interactions among the stakeholders that depend on the extant environmental features. These processes may encompass multiple layers of communicational and social factors. Language-based communication pertains to one layer of behavioural interactions and may include symbolic and non-verbal strategies. The need for communication and - more specifically - the need to understand and to predict others' intentions shape the way aesthetic decisions are made (see component 1 of the ubimus definition). Despite being among the most important achievements of the species, symbolic communication is just one way to exchange information while pursuing creative goals. Żebrowska (36), for example, points to the use of complex linguistic expressions involving graphical, mimetic, gestural or tactile components in current literary practices, "in literature, the term multimodal text (*Multimodale Texte*) can be described as what I call a multimodal message *(Multimodale Kommunikate)*." (36, page 1). Thus, although literary practices are usually centered on linguistic tokens, most of the time these tokens are accompanied by elements in other modalities.

So how is this relevant to creative music making? Ecologically grounded creative practices target a tight relationship between actions, materials and the scaffolding of the local environment. Rather than dealing with isolated sound objects - as proposed by the current Schaefferian compositional approaches - and instead of adopting the acoustic instrument, the orchestra (32), or the instrumental virtuoso (34) as the ideal models for technological development, eco-based practices embrace multimodality (1; 22), accessibility and sustainability as relevant research goals. This paper addresses two aspects of the application of eco-based creative music making. On one hand, we document the compositional strategies employed to deal with fixed imagetic material, without imposing unnecessary constraints on the shared aesthetic decisions. On the other hand, we engage with the conceptual issues that emerge from the application of eco-grounded methods in ubiquitous music practices paying particular attention to the requirements of musical collaboration in multimodal creative contexts.

2 *Atravessamentos:* Video Resources

The video material lasts six minutes and forty-seven seconds. It features shots taken in the Western Amazon rainforest, specifically in the state of Acre. Built around the participation of a dancer[4], the scenes have varied dynamics - sometimes featuring contrasting elements and other times dealing with static scenarios. In this section we take a closer look at the video[5], highlighting its potential to support the sonic aesthetic decisions (see the footage provided as supplementary material of this paper).

[4] The third author of this paper.

[5] The time stamps correspond to the fourth draft of the piece. Rather than working as a documentation of the final product, this discussion highlights the decisions involved in the creative process. Hence, we propose a snapshot of the piece taken before reaching final consensual decisions.

From the beginning of the video to minute 0:27[6], the screen remains completely dark (scene 1). The footage lasting from minute 0:27 to 1:04 features a dancer lying on dry leaves, wrapped in a veil, amidst a dark forest (scene 2). The first interaction of the dancer with the forest is featured on scene 3, with small gestures behind the forest bush (1:04 - 1:45). From minute 1:45 until 1:58 (scene 4), the dancer's face is hidden by plants and shadows. Scene 5 (1:59 - 3:01) highlights the dancer's interactions with a tree, from different angles - a far video shot and a close shot. The dancer interacts incisively, as if wishing to get inside the tree. The footage from minute 3:02 until 3:49 (scene 6) explores extensively the use of body gestures. The camera is kept distant from the subject. From 3:50 to 4:53, body gestures become complex and accentuated (scene 7), suggesting that the dancer is no longer entangled with, but confronted by the pervasive presence of the forest. On scene 8 (4:54 - 5:18), the dancer seems to establish a symbiotic relationship with her surroundings. The section lasting from minute 5:19 to 5:52 features playful gestures (scene 9). Scene 10 (5:53 - 6:43) features a change of scenery, with the dancer standing in front of a river. The gestures seem to converge with the rhythm of the natural elements (e.g., sound of rain). The last scene hints at the dancer embracing nature: the body remains fully naked and static. Scene 11 (6:64 until the end of the footage) features a moving shot of the canopy that serves as a visual background for the video credits.

Example 1. Sequence of scenes extracted from "Atravessamentos", version 4: https://youtu.be/h2kBwokW4Fo

3 Procedures

This section contextualizes the compositional strategies used in *Atravessamentos*. While tackling the methods, we explore the conceptual meanings that permeate the construction of the soundtrack.

The video was delivered semi-finalized. Although the image contents were complete, final edits were pending, thus resulting in abrupt cuts between some of the video scenes. Consensually, it was decided that a foley soundtrack was not feasible. For instance, abrupt transitions may hinder the cohesion of the sonic content. There are also moments in which the editing does not accompany the gestural elements of the dance. Instead of treating these inconsistencies as negative factors, we decided to highlight the abrupt transitions. Thus, some of these cues serve as sonic structural marks.

Both composers engaged in a collaborative process for the elaboration of the piece. Each proposal or idea was delivered to the other stakeholder to develop the existing material through modifications, additions or removal of content. Each composer had 48 hours to produce his material and hand it over to the other participant. This procedure was iterated 6 times until the result was judged as fit for evaluation by the

[6] All time stamps are given as minutes and seconds taking as reference version 4 of *Atravessamentos*. See example 1.

video producers. After the feedback from the project producers, small temporal adjustments were necessary to align the visual and the sonic contents.

The compositional problems to be surmounted featured three aesthetic aspects: 1. Consistency with the dance movements; 2. Consistency with the video cuts and with the camera positions of the video shots; 3. Use of constraints furnished by the sonic resources. This mix of creative strategies targeted: 1. A musical piece with a simple - but not necessarily simplistic - structure; 2. The use of structural articulations based on the points-of-view adopted in the video shots, serving as pillars for the sonic processes (e.g., 1:45, close-up of the dancer´s face and 4:53, scene of the *samaúma*[7] tree); 3. The use of acoustic-instrumental synthesis techniques for the excerpts with complex choreographies (scenes starting at minute 3:50 and minute 5:52 of version 4).

4 Emergent Creative Strategies

Given the aesthetic and the methodological issues described in the previous sections, we now consider the compositional strategies that frame the aesthetic solutions proposed in version 4 of *Atravessamentos*.

Biophonic gridworks. Natural sounds that hint at environmental processes provide the starting material for many of the structural compositional decisions in *Atravessamentos*. The raw elements include recordings of rain, insects, wind among a wide variety of sources. Compositional techniques derived from concrete music (30) or from soundscape composition (31) could have been employed. These techniques include audio trimming, pitch inversions and transpositions of sound objects (acousmatic procedures) or simple layering with minimal editing of recorded sounds (soundscape procedures). Nevertheless, while the adoption of a Foley-based strategy involving a linear mapping of sound events to video cuts and dance gestures presented serious limitations (see discussion in the section Procedures), an abstract soundtrack seemed to shy away from the challenge set by a footage produced by placing a dancer in outdoor, "savage" settings[8]. We opted to lay out a gridwork of biophonic sources as a sonic organizing principle. These sources were explored and expanded through a set of ecologically grounded techniques.

Mimetic strategies. The mimetic processes are directly related to natural temporal patterns found in environmental sounds. However, they also feature specific multimodal connections that foster close associations between the imagetic content and the sonic imagery. Examples include the use of biophonic sources to hint at the forest settings, synthetic pizzicato strings to emulate rain textures, and synthetic

[7] The largest amazonian tree species.

[8] Average Brazilian urban dwellers are usually mystified by the cultural production of the Amazon region. On one hand, there is the label of marginal, low-quality output attached to anything not produced in the large urban centers (typically Rio de Janeiro or São Paulo). On the other hand, there is a tendency to classify everything as "ethnic" despite the fact that the Amazon region has a long standing tradition of scientific and cultural cutting-edge contributions.

broken-glass events synchronized with the dancer's gestures. As an extension of the cues provided by the biophonic resources, mimetic strategies provide a fertile context to work with cognitive dissonances (see below).

Ecological modeling. Since Keller and Truax (24) initial implementations, ecological models have been applied to a variety of contexts - ranging from multimodal installations (20) to multimedia projects (2; 17), audio haptics and textural synthesis. More recently, Yoganathan (35) proposes simple mixing strategies - involving the usage of geographically disparate sonic cues - to explore conflicting renditions of soundscape recordings (35), labeled *artificial ecotones*. Another thread of development is explored by (4), targeting the resynthesis of realistic sonic scenes from short samples. *Atravessamentos* makes use of ecological modelling to expand the palette of available sonic resources with close ties to the behaviours of everyday sonic events. More specifically, a contribution of this piece to current developments in ecologically grounded creative practice is the incorporation of acoustic-instrumental synthesis within the context of eco-based methods.

Acoustic-instrumental synthesis. Atravessamentos features sonic generative techniques based on acoustic instrumental sources. These emulations are produced by decoding the dynamic spectral structure of the recorded sources. These synthesis processes seek to recreate or emulate acoustic instrumental sounds, while allowing for the timbral expansion of the instrumental palette[9]. Furthermore, acoustic-instrumental recordings are modified via sound-processing methods, such as reverberation and compression. We employ audio-processing techniques to enhance the perceptual qualities of the recorded materials and to establish timbral bridges among the synthetic and the recorded resources. In *Atravessamentos*, we use instrumental samples as dynamic clusters and glissando sound blocks to yield dense sonic textures. These generative techniques target the emergence of cognitive dissonances through manipulations of the expected sonic cues of a biophonic soundscape. The expanded sonic palette is enabled by ecological modeling and acoustic-instrumental synthesis.

Cognitively dissonant strategies. For creative purposes, c*ognitive dissonances* involve the use of conflicting matches among elements of various modalities, targeting the distortion of perceptual cues and leading to a larger set of aesthetic relationships among images, sounds, behaviours and spatial and temporal cues. Both the images and the sonic materials of *Atravessamentos* feature cognitive dissonances at various levels. A key element is the dancer's immersion in the scenery of the Amazon forest. The presence of the dancer represents an unusual phenomenon in this context. This disruption suggests the use of sonic processes and sources that are not directly related to the original context, hence "breaking" or suspending the listener's expectations.

Visually, rain images are not featured until the last video scene (with the dancer at a margin of a river). The sound of rain serves as a scaffolding resource at a macro level: it is at the foreground of the initial section and it is also featured throughout the final

[9] There are several methods available for this compositional strategy, including the pioneering proposals by J. C. Risset - analysis and synthesis based on Fourier models - and more recent approaches, such as spectral modeling, group additive synthesis and analysis and transformation synthesis.

section of the work. At a meso level[10], it serves as a "percussive temporal grid" to set the timings of some of the most salient sonic events. The rain texture also furnishes a timbral bridge between the biophonic and the synthetic materials. Given that the string pizzicatos and the synthesized glass sounds are sonic classes of highly unlikely occurrence within the context of the Amazonian soundscape, they become good candidates to trigger cognitive dissonances. However, to achieve the intended perceptual paradoxes it is necessary to keep a consistent set of geo-location references (see 19 for previous endeavors that employ this approach). These local referents - or anchors - are established through the use of the Amazonian forest soundscape recordings.

Textural density. This organizing principle is applied to various types of materials - such as the biophonic sources, the synthesised sources and the sound processing of acoustic or electronic instrumental sources. The ability to increase or to reduce the density of sonic events may enhance the emotional impact of the scenes. Thus, changes in sound density in *Atravessamentos* are closely related to the erotic content of the images, seeking a connection between the dancer's gestures and the aesthetic usage of sonic cues.

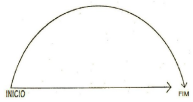
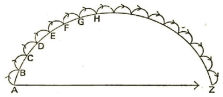
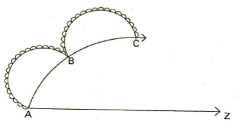
5 Contributions and Implications for Artistic Practices

This work brings several contributions to musical creative practices based on ecological cognition. The soundtrack combines Western Amazonian sonic sources, acoustic-instrumental synthesis and ecological modelling with videodance materials collected in the Western Amazon. The compositional strategies are aligned with previous proposals by (1; 2; 16; 17; 19). However, this is the first work that unites acoustic-instrumental synthesis with ecological modelling of everyday sounds. Furthermore, while cognitive dissonances are explored by other authors, to the best of our knowledge no attempt has been made at grounding this method on a firm conceptual basis. So let us elaborate this point in more detail.

As discussed in the section Related Approaches, exploratory creative processes that lead to alternative aesthetic decisions are usually defined as *detouring* or *shifting* (11; 26). According to Mannis (26), shifting entails an unexpected chain of events within the creative process. The alternative paths yielded by shifting may be resolved as a definite musical structure. Other times they remain unresolved. Esslin's (11) diagram illustrates how the aesthetic decision processes can take place either linearly or through multilevel stages (table 1). According to Esslin (11), shifting may be employed for dramatic purposes - targeting an increase of the emotional load to maintain the attention of the listener. We propose that instances of this strategy are exemplified by the methods that target cognitive dissonances.

[10] See discussions on micro, meso and macro compositional strategies in (16; 19).

Table 1. Creative strategies according to Esslin (1986).

Structure	Diagram
Linear chaining. Linear dramatic construction suitable for cinematographic development (10; 26).	
Recursive process targeting sustained attention.	
Recursive chaining.	

The process of shifting in *Atravessamentos* encompasses several aspects. For the suspension of expectations to occur, it is necessary to keep a congruent set of elements aligned with the contextual normative aspects. The *biophonic gridworks* provide triggers for associations with the Amazonian forest settings. Sonicly, these expectations are broken through the introduction of acoustic-instrumental events, though this resource is used subtly not to obliterate the sonic anchors. The application of shifting targets building and breaking expectations of recognizable multimodal properties, rather than establishing a strictly causal sequence of events. Therefore, we also avoid employing linear chaining. Furthermore, we address the need of expanded alternatives to the use of explicit knowledge fostered by the teleological approaches that target problem solving. For instance, Aliel et al. (2) propose the adoption of meditative thought as an application of Heidegger´s *Gelassenheit* concept within the context of ecologically grounded creative practice. Their usage of meditative thought is tailored as an adaptation of the agents during the creative act, not necessarily involving a change of perspective. In *Atravessamentos* the meditative approach is applied as a creative strategy that entails a breach of expectations. As a consequence of unexpected contents and behaviours, the participant is encouraged to change her worldviews triggering a process that may lead to a paradigm shift.

6 Acknowledgements

This project was partially funded by a CNPq Productivity Research Grant [300996/2018-7] to Damián Keller. His participation at the UbiMus 2019 Workshop was made possible by the CNPq [450755/2019-3] and by the organizing committee of the CMMR 2019. We would like to thank Mathieu Barthet and Leandro Costalonga for their assistance.

7 References

1. Aliel, L.; Keller, D. & Costa, R.: Comprovisation: An approach from aesthetic heuristics in ecocomposition (Comprovisação: Abordagens desde a heurística estética em ecocomposição). In: *Proceedings of the Brazilian Symposium on Computer Music (SBCM 2015)*, 169-180. Campinas, SP: SBC (2015).
2. Aliel, L.; Keller, D. & Costa, R.: Theoretical perspectives for the analysis of ecologically grounded creative practice (Perspectivas teóricas para a análise das práticas criativas ecocognitivas). In: Damián Keller, Helena Lima (eds.), *Ubiquitous Music Applications* (*Aplicações em Música Ubíqua).* São Paulo, SP: ANPPOM (2018).
3. Barrass, S.: The musification of furniture in the form of a pouf-doodle. In: *Proceedings of the VI Workshop on Ubiquitous Music (VI UbiMus)*. Växjö, Sweden: Ubiquitous Music Group (2015).
4. Bernardes, G.; Aly, L. & Davies, M. E. P.: SEED: Resynthesizing environmental sounds from examples. In: *Proceedings of the SMC (SMC 2016)*. Hamburg, Germany: SMC (2016).
5. Bown, O.; Eldridge, A. & McCormack, J.: Understanding interaction in contemporary digital music: From instruments to behavioural objects. *Organised Sound* 14 (2), 188-196 (2009).
6. Capasso, A.; Keller, D. & Tinajero, P.: Sisyphus/Sísifo 1.0 [Multimedia Installation]. New York, NY: General Consulate of Argentina (2004).
7. Donald, M.: Art and cognitive evolution. In: Turner, M. (ed.): *The Artful Mind* (pp. 3-20). Oxford, UK: Oxford University Press (2006).
8. Dunbar, R. I. M. & Shultz, S.: Evolution in the social brain. *Science* 317 (5843), 1344-1347 (2007).
9. Einarsson, A. & Ziemke, T.: Exploring the multi-layered affordances of composing and performing interactive music with responsive technologies. *Frontiers in Psychology* 8, 1701 (2017).
10. Emmerson, S.: The relationship of language to materials. In: Emmerson, S. (ed.), *The Language of Electroacoustic Music* (pp. 17-39). London, UK: Palgrave Macmillan (1986).
11. Esslin, M.: An anatomy of drama (Uma anatomia do drama). Rio de Janeiro, RJ: Zahar (1986).
12. Ferraz, S. & Keller, D.: Preliminary proposal of the In-group, Out-group model of collective creation (MDF: Proposta preliminar do modelo dentro-fora de criação coletiva). In: *Cadernos de Informática* 8 (2), 57-67 (2014).
13. Gibson, J. J.: The theory of affordances. In: Shaw, R. & Bransford, J. (eds.): *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (pp. 67-82). Mahwah, NJ: Lawrence Erlbaum Associates (1977).

14. Gibson, J. J.: *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin (1979).
15. Hutchins, E.: Cognitive ecology. *Topics in Cognitive Science* 2 (4), 705-715 (2010).
16. Jones, D.; Brown, A. R. & d'Inverno, M.: The extended composer. In: McCormack, J. & d'Inverno, M. (eds.): *Computers and Creativity* (pp. 175-203). Berlin and Heidelberg: Springer (2012).
17. Keller, D.: Compositional processes from an ecological perspective. *Leonardo Music Journal* 10(5), 55-60 (2000).
18. Keller, D.: Sonic Ecologies. In: Brown, A. R. (ed.): *Sound Musicianship: Understanding the Crafts of Music* (pp. 213-227). Newcastle upon Tyne, UK: Cambridge Scholars Publishing (2012).
19. Keller, D.; Barreiro, D. L.; Queiroz, M. & Pimenta, M. S.: Anchoring in ubiquitous musical activities. In: *Proceedings of the International Computer Music Conference* (pp. 319-326). Ann Arbor, MI: MPublishing, University of Michigan Library (2010).
20. Keller, D. & Capasso, A.: New concepts and techniques in eco-composition. In: *Organised Sound* 11 (1), 55-62 (2006).
21. Keller, D. & Lazzarini, V. Ecologically grounded creative practices in ubiquitous music. Organised Sound 22 (1): 61–72, 2017.
22. Keller, D.; Lazzarini, V. & Pimenta, M. S. (eds.): *Ubiquitous Musi*c, XXVIII. Berlin and Heidelberg: Springer International Publishing (2014a).
23. Keller, D.; Otero, N.; Lazzarini, V.; Pimenta, M. S.; Lima, M. H.; Johann, M. & Costalonga, L.: Relational properties in interaction aesthetics: The ubiquitous music turn. In: Ng, K.; Bowen, J. P. & McDaid, S. (eds.): *Proceedings of the Electronic Visualisation and the Arts Conference (EVA 2014)*. London: BCS, Computer Arts Society Specialist Group (2014b).
24. Keller, D. & Truax, B.: Ecologically based granular synthesis. In:: *Proceedings of the International Computer Music Conference* (pp. 117-120). Ann Arbor, MI: MPublishing, University of Michigan Library (1998).
25. Lazzarini, V.; Keller, D.; Kuhn, C.; Pimenta, M. & Timoney, J.: Prototyping of ubiquitous music ecosystems. In: *Journal of Cases on Information Technology* 17, 73-85 (2015).
26. Mannis, J. A.: Processos cognitivos de percepção, análise e síntese atuando no processo criativo: Mímesis de mímesis. In:: *Anais do Encontro Nacional de Composição Musical de Londrina (EnCom 2014)*. Londrina, PR: UEL (2014).
27. Nance, R. W.: *Compositional explorations of plastic sound*. Doctoral Thesis in Music. De Montfort University, UK, 2007
28. Odling-Smee, F. J.; Laland, K. N. & Feldman, M. W.: *Niche Construction: The Neglected Process in Evolution*. Princeton, NJ: Princeton University Press (2003).
29. Rhodes, M.: An analysis of creativity. *The Phi Delta Kappan* 42, 305-311 (1961).
30. Schaeffer, P.: *Traité des objets musicaux: Essai interdisciplines*. Paris: Éditions du Seuil (1966).
31. Schafer, R. M.: *The Tuning of the World*. New York, NY: Knopf (1977).
32. Trueman, D.: Why a laptop orchestra?. *Organised Sound* 12 (2), 171-179 (2007);
33. Turchet, L.; Fischione, C.; Essl, G.; Keller, D. & Barthet, M.: Internet of Musical Things: Vision and Challenges. *IEEE Access* 6, 61994-62017 (2018).
34. Wessel, D. & Wright, M.: Problems and prospects for intimate musical control of computers. *Computer Music Journal* 26 (3), 11-22 (2002).
35. Yoganathan, N.: *Disparate soundscapes and ecotones: Critically sounding the Amazon and Arctic*. Master of Arts in Media Studies. Montreal: Concordia University (2017).
36. Żebrowska, E.: Multimodal messages. *Journal of Multimodal Communication Studies* 1, paper 5 (2011).

The Analogue Computer as a Voltage-Controlled Synthesiser

Victor Lazzarini and Joseph Timoney

Maynooth University,
Maynooth, Co. Kildare,
Ireland
{victor.lazzarini,joseph.timoney}@mu.ie

Abstract. This paper re-appraises the role of analogue computers within electronic and computer music and provides some pointers to future areas of research. It begins by introducing the idea of analogue computing and placing in the context of sound and music applications. This is followed by a brief examination of the classic constituents of an analogue computer, contrasting these with the typical modular voltage-controlled synthesiser. Two examples are presented, leading to a discussion on some parallels between these two technologies. This is followed by an examination of the current state-of-the-art in analogue computation and its prospects for applications in computer and electronic music.

Keywords: analogue computing, voltage control, sound synthesis, filters, oscillators, amplifiers, FPAA, VLSI analogue circuits

1 Introduction

Computer Music, for the most of its history, has been concerned with the use of what we can generally class as the *digital stored-program computer*, although this (correct) terminology has by now fallen in disuse, due to the ubiquitous nature of these devices. In this paper, we will instead look at a different type of computer and its potential to sound and music design, and its relationship to the music instrument technology of voltage control. The principles we will explore fall into the category of *analogue computing*, which approaches both the actions involved in computation, the modelling and the problem design from a different perspective.

The principles that constitute analogue computing can be seen from two perspectives that are somewhat independent from each other. On one hand, the hardware that implements it allows for the solution of problems containing continuous variables, whereas digital circuitry implies a discretisation of these (even in models that assume underlying continuous quantities). In the case of music, the possibility of time and amplitude-continuous computation is significant, considering the amount of work that has been dedicated to solving discretisation issues in areas such as virtual analogue models [17].

From a different perspective, analogue computing approaches problems in a way that largely dispenses the algorithmic approach of digital computer programming in favour of hardware reconfiguration, setting up the computation not so much as a sequence of steps, but as an interconnection of components [22]. This also implies that the hardware model set up in this way is an *analogue* of the problem at hand. Additionally, while analogue computers may be able to compute problems related to the steady state of a system, they are more frequently used to providing solutions relating to transient behaviour [14]. Such problems are significant to sound and music applications, where the dynamic properties of a system are fundamental.

Analogue computing has had a long history, which began with mechanical devices that were used as aids to calculation of specific problems (navigation, gunnery, accounting, etc.), and became a major scientific field of research with the advent of practical electronic devices. These could be combined more flexibly to realise various types of modelling. From a music perspective, these developments influenced the technology of voltage control, and the modular aspect of electronic analogue computers appears to be significant in providing the priciples underpinning early synthesisers [12].

In this paper, we will examine the relationships that exist between these devices. We will start by exploring the principles of analogue computation with electronic computers. This will be followed by an introduction to modular voltage-controlled synthesisers from the perspective of analogue computing. Then we will examine the possibilities of general-purpose electronic computers as musical instruments, followed by an examination of the current state of the art in the area and the perspectives for new research in sound and music (analogue) computing.

2 Electronic Computers

Analogue computers, as discussed in the introduction to this paper, operate under different principles to their digital stored-program counterparts. Generally, they are set up to provide solutions to a problem that is laid out in terms of a mathematical equation or set of equations, providing an answer to these, given a certain input. In this case, the type of problems that are applied to them can be of different characteristics, provided that they can be described in an algebraic form. Programming the computer is then a matter of setting an analogue to the original problem [22] by means of various computing elements. Therefore, the capabilities of a given analogue computer are determined by the types of computing blocks it can offer, and how they can be connected in a program.

2.1 Computing elements

Analogue computers are made of various types of components that often operate as *black boxes*, providing an output given a set of inputs and conditions, within a certain level of tolerance. The inputs and outpus of such boxes are electric signals whose voltages play the part of the variables that are manipulated in a

program. Programs will then be made up of patching connections between these different blocks, setting up the initial conditions that configure the problem and then running the computer, from which the answer or answers can be read by appropriate output devices.

While the components of an analogue computer can be quite varied in nature, there are some key blocks that are present universally in these devices, to provide basic computing operations.

Arithmetics We can divide the arithmetic operations into three fundamental categories, that are addressed by specific types of electronic circuits: (a) multiplication by a scalar; (b) addition/sum; (c) multiplication of signals. In the case of (a) and (b), a fundamental component is the *operational amplifier* [19]. This component allows a gain to be applied to the signal, and facilitates both multiplication and addition to be implemented. Of course, if only attenuation is required, then a signal can be passively modified by a variable resistance (fig. 1), but in all other cases, the op amp is required.

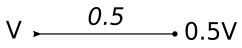


Fig. 1. Attenuation example

Gain scaling is implemented simply by setting the multiplier constant k in the op amp, which is the ratio of the resistances R_f/R_i that are employed in the circuit (fig. 2)

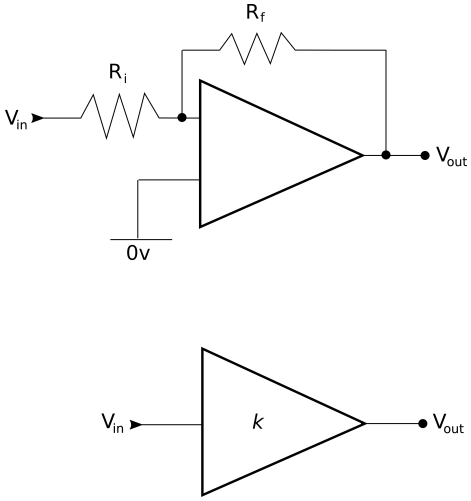


Fig. 2. Op amp circuit schematics and gain scaling symbol

$$V_{out}(t) = -kV_{in}(t) \tag{1}$$

Note that the op amp will normally have the effect of inverting the sign of the voltages applied to its input, due to the fact that only its inverting input is used.

Summing two voltages also require an op amp (fig. 3), and the input signals are scaled by the ratios of the individual input resistances and the feedback path resistance, $k = R_f/R_n$. Note that adding units such as these can be set up for more than two inputs.

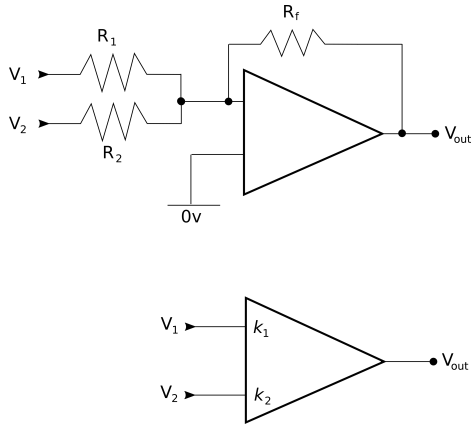


Fig. 3. Summing amp circuit with two inputs and symbol

$$V_{out}(t) = -\sum_{0}^{n} k_n V_n(t) \tag{2}$$

Multiplication of two signals is generally taken as a separate category as it requires more complex circuitry. In this case, the output is equivalent to the instantaneous value of the multiplication of its inputs, scaled by a constant.

Integration Another key component of an analogue computer is the integrator. The means of integrating an input signal is provided by a capacitor, and the circuit (fig. 4) also includes an op amp to complement it. As we can see, the capacitor replaces the feedback resistor in a simple scalar multiplier. The output is also scaled by $k = 1/R_iC$ where C is the capacitance in the op amp feedback path. The voltage across the capacitor can also be set as an initial condition V_0.

$$V_{out}(t) = -k \int_0^t V_{in}(t) + V_0 \tag{3}$$

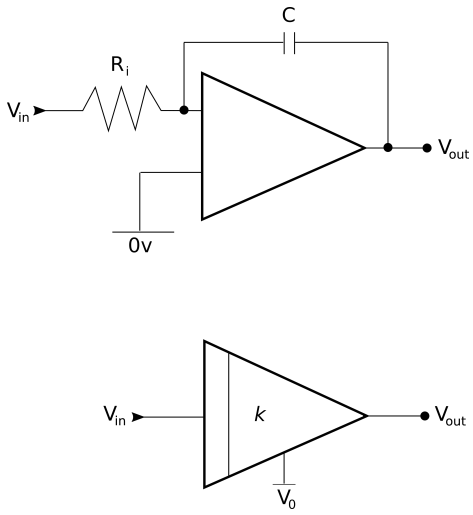


Fig. 4. Integrator circuit and symbol

It is also a simple matter to include multiple input signals to an integrator, using a combination of the circuits of figs. 2 and 3. In this case, the different inputs are scaled and added together before the integration is performed.

Functions It is also fundamental for analogue computers to be able to provide means of generating a variety of functions. Among these we will find the usual single-variable functions trigonometric, exponential, triangle, rectangular, ramp, etc. Some computers would also have more sophisticated means of generating user-defined functions [22]. It is worth noting that function generators is a general class of modules that also include the multiplication, summation, and integration blocks described above [14].

Other Modules Various other modules exist in various analogue computing devices [22]. Logic blocks such as comparators allow voltages to be compared for binary decisions (such as opening and closing signal connections) and step-function implementations. Limiters are special-purpose comparators that can keep signals within a given range. Time delays provide a means of shifting the phase of functions and can be implemented in discrete capacitor circuits called bucket brigade devices [20] Output of analogue computations involve some means of measuring the voltage of a program, which can be done by various means such as strip-chart and *xy* recorders, oscilloscopes, voltmeters, and similar components. A sound and music computing relevant output block would consist of an audio pre-amplifier that allows line-level connections to a mixer and power amplifier. This is of course, the main output component of a voltage-controlled synthesiser.

3 Modular Synthesisers

Modular electronic sound synthesis, especially the variety involving voltage control technologies, has been a cornerstone of electronic music since the post-war era [23]. In examining the devices that have been and are currently used for this purpose, we can see very clear parallels with the technology of analogue computers. In fact, analogue synthesisers in general have been identified as special purpose computers [22]. As in that type of technology, modules play an important part as the components of programs, which in the case of the typical modular design are made up of patch-cord connections between them.

3.1 Modules

Voltage-controlled synthesizer modules are generally built at a higher level of operation if compared to analogue computing blocks. This means that access to fundamental aspects of computation are less common. For example, function generators in the form of oscillators realise compound operations, which in the case of an analogue computer would be provided in smaller building blocks of a time function generator, plus multipliers and adders. However, some basic elements are given: ring modulators implement two-input multiplication, mixers are summing modules, and offset/scaling modules can provide addition and multiplication of signals and scalars. The typical modules of a synthesiser include voltage controlled filters (VCFs), oscillators (VCOs), and amplifiers (VCAs).

In general, modular synthesisers provide a rich set of components, many of which can be seen as different types of function generators: for attack-decay-sustain-release curves, noise sources, and sequencers, which provide user-defined functions. However, the synthesiser set of components is provided with significant specialisation, and in general lacks access to the fundamental building blocks of computation. Tracing an analogy to the music languages used for programming digital computers, modular synthesizers provide high-level unit generators [6], but not the means of coding (or in this case, setting up) the unit generators themselves (the modules in the analogue domain).

4 Examples and Discussion

In order to the expand the discussion of analogue computing for sound and music, it is interesting to consider some examples to illustrate simple operations. While we would put these problems from a general-purpose computing perspective, we would also like to consider them with respect to typical sound synthesis applications.

4.1 Linear Functions

The simplest example of the application of analogue computation is to set up the solution to a linear problem, such as

$$f(t) = ax(t) + b \tag{4}$$

which may be applied, for instance, to glide the pitch of a tone from one frequency to another. The program for this is shown in figure 5, smoothly sliding by a user-defined interval. In this case, each increment of 1V in the input starting from a voltage V_0, will provide a jump of k semitones, when used as a 1V/oct exponential frequency signal.

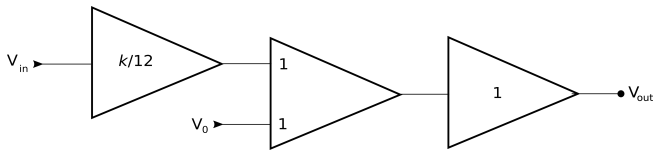


Fig. 5. Linear equation program

This, of course, can easily be set up in a modular synthesiser by the use of an amplifier and an offset, which are blocks that are readily available. At this level of simplicity, the synthesiser can match the analogue computer almost on a one-to-one component basis.

4.2 Differential Equations

A more common application of analogue computers has to do with the solution of differential equations. Consider the following example,

$$y(t) = ax(t) - b\frac{dy(t)}{dt} \tag{5}$$

which is a simple first-order differential equation. This can be translated into an analogue computer program as shown in fig. 6. The significance of this is that such a differential equation also implements a simple infinite impulse response low-pass filter. With this approach, could implement filters of different designs, more complex of course, but using the common blocks of an analogue computer.

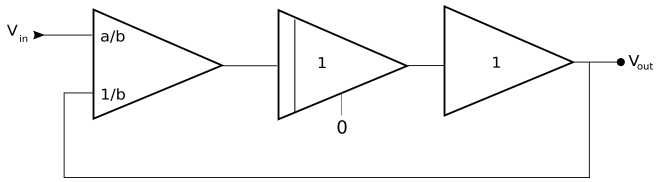


Fig. 6. Difference equation program

Hutchins [4] pointed out that the fundamental voltage-controlled building block used here (fig. 6), the integrator, is applicable to the well-known state variable filter (SVF) design. He noted that the fact that the SVF is formed from elemental blocks of two integrators and a summer in a loop has been known from analogue computer programs, where it was used in physical simulations of second-order responses by producing voltages corresponding to the magnitudes of the *state variables* of the system.

This first-principles approach is contrasted now with the filter modules implementing different topologies, with various particular characteristics, that are found in voltage controlled synthesisers. The significant difference is that these are fixed to a given design, and do not allow access or manipulation of its circuit connections. This demonstrates an example where there is no one-to-one match between an analogue computer and a synthesiser.

4.3 The General-purpose Analogue Computer as a Musical Instrument

Given the examples discussed above, it might be surprising to see that not a lot has been made in terms of utilising general-purpose analogue computers in musical applications. The composer Hans Kulk appears to be a solitary figure working in this field [5]. This may be attributed to various factors: cost, as in the heyday of analogue computers, it was very expensive to access these; complexity, programming analogue computers required significant technical expertise, which was not mitigated by music-directed approaches, as for instance provided by music programming languages in the case of digital computers.

The existence of the modular synthesiser, in fact, can be seen as the analogue counterpart to the digital computer music programming environments. However, as noted above, they were not developed to provide lower-level access to computation, which is the case of digital computer programming (as in, for instance, Csound [7]. Finally, also we should consider that the obsolescence of the general-purpose analogue computer, in parallel with the ubiquitousness of the digital computer, also played a part in the process. However, some new prospects in the analogue computing domain may allow us to re-appraise these devices as possible vehicles for music making. The main consideration is that there is no fundamental impediment to this; on the contrary, there seems to be fertile ground for work in this respect.

5 Prospects for Electronic and Computer Music

While analogue computers may appear to some to have passed their heyday and be generally obsolete today, that does not seem to be the case if we look at some cutting-edge research in the field. The direction of travel seems to involve the development of very large scale integration (VLSI) components that implement a collection of computing elements. These can then be made available for programming in flexible ways and used as dedicated processors within a larger

host system [2]. The technology of Field Programmable Analogue Arrays [9–11] has also opened some interesting possibilities in the particular case of analogue signal processing.

It is within this context that a number of prospects for musical signal processing may arise. The interest in re-creating analogue environments using digital simulations that sparked virtual analogue model research may have exhausted its potential for further refinements, and work that is actually directed to the *real* thing is an attractive proposition. Music production is one of the few areas of technology where analogue signal generation and processing techniques continue to be used, and in fact, have enjoyed a significant resurgence

5.1 Field Programmable Analogue Arrays

Recent developments in analogue signal processing technology included the development of Application-specific Integrated Circuits (ASICs), used in the implementation of synthesiser modules such as filters and oscillators. However, as the name implies, ASICs target specific purposes and re-designing them is very expensive. What would be more suitable is to have an analogue equivalent of the digital Field Programmable Gate Array (FPGA). Fortunately, the concept of Field Programmable analogue Arrays (FPAAs) was introduced with the promise that it facilitated analogue components to be connected together in an arbitrary fashion, allowing for rapid testing and measurement of many different circuit designs.

A similar but less sophisticated technology is the PSoC (programmable system-on-chip) by Cypress Semiconductor. These chips include a CPU core and mixed-signal arrays of configurable integrated analogue and digital peripherals. The FPAA was introduced in 1991 by Lee and Gulak [9]. The idea was further enhanced by the same authors in 1992 [10] and 1995 [11] where op-amps, capacitors, and resistors could be connected to form a biquad filter, for example. In 1995, a similar idea, the electronically-programmable analogue circuit (EPAC) was presented in [18].

Within the FPAA explored in [15] it was organised into three functional blocks: (1) the computational analogue block (CAB), which is a physical grouping of analogue circuits that act as computational elements, (2) the switch matrix (SM), which defines the interconnection of CAB components, and (3) the programmer which allows each device to be turned completely on, turned completely off, or operated somewhere in-between. This flexibility means that switch elements can be used for computation as well as routing. This is especially beneficial in audio applications, since transistors set to a constant bias are often necessary.

Only in recent years have FPAAs become powerful enough to be considered for facilitating complex analogue sound synthesis, in the implementation of common modules. Two papers investigated whether FPAAs were capable of creating entire synthesis systems. One paper illustrated how the low-pass VCF developed and popularised by Robert Moog [13] could be implemented using an FPAA [15]. For this implementation it was found that the FPAA could support 12 VCFs,

assuming perfect utilisation of all the available resources by the CAB, but it may also be possible to include another 8-10 filters under alternative constraints. A second paper looked at the FPAA configuration for a VCO and VCA, and two common control modules, the low-frequency oscillators and the envelope generator, which would allow for the development of a complete synthesiser [16]. The paper identified a number of challenges, which included whether the VCO implementation would be controllable over a wide pitch range and remain stable with temperature changes. In general, FPAAs appear to be one of the most promising analogue technologies for sound and music computing.

5.2 Programmability

Programming a modular synthesiser with patch-cords is a significant undertaking, and the traditional analogue computers presented much bigger challenges in that respect. Clearly, if we are to be able to implement signal-processing operations from first principles, the question of programmability takes is a key concern. In modern VLSI-based systems such as the one described in [2] and in the case of FPAAs, some means of setting up connections between components is provided (and storing/retrieving these), which can be at various levels.

We could trace a parallel with digital computers, where the code may be represented by an assembly-type language for a given hardware, or in a high-level language such as C. In an analogue computer, we could manually translate equations (such as for instance eq. 5) into the actual hardware connections (e.g. fig. 6), or potentially use an algebraic compiler to synthesise the necessary circuits (such as the one discussed in [1]). We can hypothesise that such a high-level language could be developed targeting the requirements of analogue signal processing for music computing applications.

5.3 Hybrid Digital-Analogue Systems

Another emerging characteristic of modern analogue computing appears to be the development of hybrid digital-analogue systems. One such arrangement is described in [3], where a combination of digital and analogue circuits are used to construct a programmable device. This type of arrangement is mirrored in modern polyphonic analogue synthesisers, where audio signals are kept in the analogue domain, and control signals originate from digital representations and are transformed into voltage control via a number of digital-to-analogue converters. This allows some level of interconnectivity via so-called modulation matrices that mimic the modular approach, albeit in a smaller scale.

6 Opportunities for Ubimus

Within the perspective of ubimus, technologies such as the ones discussed here can open up new possibilities for approaches that aim to reproduce early electronic music practices. From the perspective of programmability, which is one of

the cornerstones of ubimus practices [8], the aim is to provide musicians of all levels, as well as researchers, flexible tools that would allow them to manipulate sound with analogue means. The technology is not prescriptive in terms of what its applications are, and that open-endedness can be translated into components of digital-analogue devices for use in, for instance, IoMusT [21] applications in professional and everyday-music settings.

7 Conclusions

This paper attempted to demonstrate the usefulness of an analogue computing approach to electronic and computer music research. It provided a general introduction to the area, alongside tracing a parallel to modular voltage-controlled synthesisers. Examples were given that had direct relevance to analogue audio signal processing, demonstrating some immediate applications in research and music production.

A survey of the state-of-the-art in analogue computing provided us with some first candidates as technologies that might be ready for use. In fact, in one case some interesting results had already been presented. Challenges remain, however, given that the target outputs of analogue computing for music applications have some key constraints of quality, including low signal-to-noise ratios and pitch/voltage stability. Assessing this should play an important part in any future research.

Another aspect that was raised was to do with programmability. We see that key developments in the area are necessarily linked to the potential of music programming systems. Having analogue computing-dedicated music tools, similar to what exists for standard digital computers, will play an important part of making the technology available to a wider range of users (musicians, artists). This possibly points out to another fertile field of computer music research.

References

1. Achour, S., Sarpeshkar, R., and Rinard, M. Configuration synthesis for programmable analogue devices with Arco. In *Proceedings of PLDI 16* (Santa Barbara, CA, 2016), pp. 177–193.
2. Cowan, G., Melville, R., and Tsividis, Y. A VLSI analogue computer / digital computer accelerator. *Journal of Solid-State Circuits 41*, 1 (2006), 42–53.
3. Guo, N., Huang, Y., Mai, T., Patil, S., Cao, C., Seok, M., Sethumadhavan, S., and Tsividis, Y. Energy-efficient hybrid analog/digital approximate computation in continuous time. *Journal of Solid-State Circuits 51*, 7 (2016), 1514–1524.
4. Hutchins, B. Revisiting some vcf ideas – and a few new ideas. *Electronotes 23*, 215 (2013), 1–23.
5. Kulk, H. Proposal for extending analogue modular electronic music synthesizers with function modules from the analogue computation repertoire. In *Symposium Think Analogue! Humboldt University* (Berlin, 2012).
6. Lazzarini, V. The development of computer music programming systems. *Journal of New Music Research*, 42 (2013), 97–110.

7. Lazzarini, V., ffitch, J., Yi, S., Heintz, J., Brandtsegg, Ø., and McCurdy, I. *Csound: A Sound and Music Computing System*. Springer Verlag, 2016.
8. Keller, D., Lazzarini, V., and Pimenta, M. *Ubiquitous Music*. Springer Verlag, 2014.
9. Lee, E. K. F., and Gulak, P. G. A cmos field-programmable analog array. *IEEE Journal of Solid-State Circuits 26*, 12 (Dec 1991), 1860–1867.
10. Lee, E. K. F., and Gulak, P. G. Field programmable analog array based on mosfet transconductors. *Electronics Letters 28*, 1 (Jan 1992), 28–29.
11. Lee, E. K. F., and Gulak, P. G. A transconductor-based field-programmable analog array. In *Proceedings ISSCC '95 - International Solid-State Circuits Conference* (Feb 1995), pp. 198–199.
12. Moog, R. Voltage controlled electronic music modules. *AES Preprint 346* (1967), 1–19.
13. Moog, R. Electronic high-pass and low-pass filters employing the base to emitter diode resistance of bipolar transistors, 1969. US Patent 3475623A.
14. Navarro, S. *analogue Computer Fundamentals*. Wadsworth Publ. Co., Belmont, CA, 1962.
15. Nease, S. H., Lanterman, A. D., and Hasler, J. O. A transistor ladder voltage-controlled filter implemented on a field programmable analog array. *J. Audio Eng. Soc 62*, 9 (2014), 611–618.
16. Nease, S. H., Lanterman, A. D., and Hasler, J. O. Applications of current-starved inverters to music synthesis on field programmable analogue arrays. *J. Audio Eng. Soc 66*, 1/2 (2018), 71–79.
17. Pakarinen, J., Valimaki, V., Fontana, F., Lazzarini, V., and Abel, J. Recent advances in real-time musical effects, synthesis, and virtual analogue models. *Eurasip Journal On Advances In Signal Processing 2011:940784* (2011), 1–15.
18. Pierzchala, E., Perkowski, M. A., Van Halen, P., and Schaumann, R. Current-mode amplifier/integrator for a field-programmable analog array. In *Proceedings ISSCC '95 - International Solid-State Circuits Conference* (Feb 1995), pp. 196–197.
19. Ragazzini, J., Randall, R., and Russell, F. Analysis of problems in dynamics by electronic circuits. *Proceedings of the IRE 35* (1948), 444–452.
20. Sangster, F., and Teer, K. Bucket-brigade electronics - new possibilities for delay, time-axis conversion, and scanning. *IEEE Journal of Solid State Circuits 4*, 3 (1969), 131–136.
21. Turchet, L., Fischione, C., Essl, G., Keller, D., and Barthet, M. Internet of Musical Things: Vision and Challenges. *IEEE Access 6*, (2018), 61994-62017.
22. Ulmann, B. *Analog Computing*. Oldenbourg, Munich, 2013.
23. Wells, T. *The Technique of Electronic Music*. Schirmer Books, New York, 1981.

Sounding Spaces for Ubiquitous Music Creation

Marcella Mandanici

Music Conservatory "Luca Marenzio"
Brescia (Italy)
mmandanici@gmail.com

Abstract. This paper discusses the use of large-scale responsive environments for ubiquitous music creation. A short analysis of how people listen to music in everyday life settings and of active listening practices and applications outlines the function of everyday music listening and the shift towards the concept of ubiquitous music creation. Three models for the creative use of space are presented based on music application examples: the bi-dimensional slider, the interactive landmarks and the grid. Large-scale interactive environments may be embedded in public and private spaces and allow music creative practices for many users, learners or people with disabilities.

Keywords: Large-scale interactive environments, full-body interaction, active listening, ubiquitous music creation

1 Introduction

In the context of this contribution the author defines large-scale responsive environments as floor surfaces where the users presence, motion and gestures can be tracked by a computer system equipped with cameras or/and motion sensors. Computer vision algorithms process the incoming data by producing the coordinates of the users position and movements, which can be employed to provide an audio output coherent with the users actions. This builds a strong relationship between the position or movements of the users and the environment around them. Starting from these characteristics the author has experimented with various music applications which share some common principles with *ubiquitous music* research. The main one is the search for musical production methods alternative to the use of traditional musical or electronically augmented instruments. As pinpointed by Keller et al. [1] there are many open issues in current musical practices such as the need for concert halls specifically dedicated to musical activities, the limitations due to the difficulty of instrumental techniques, and the clear separation of roles between performers and public. These issues represent important drawbacks not only in the field of music creation but also for music education and for the inclusion of children with various disabilities [2]. Large-scale responsive environments partially respond to these questions. They well fit Mark Weiser's vision of ubiquitous computing, where technology is embedded in every day objects but, at the same time, is not visible to the user [3]. As

true, physic spaces which can be positioned in public venues such as classrooms, libraries, museums, as well as in private homes, they allow music production in every kind of public of private setting. Moreover, the unencumbered, completely natural interaction style proper of large-scale responsive environments fully resonates with Mark Weiser's idea of unobtrusive, simple technology. In fact users can make music through the simple act of entering the environment's active area, without the need of learning particular instrumental techniques. But, whereas ubiquitous computing is based on the presence of smart objects scattered in the surroundings, large-scale responsive environments, on the contrary, claim for a strong, localized presence, or even for a social role in the community [4]. Like a modern *agora* these environments can attract the attention of people in the surroundings while some others enter the active area. Thus music creation becomes a social event where the user's experience is public and open to comments and suggestions of bystanders. This can happen in gaming [5], entertainment [6], therapy [7] and learning [8], [9] and [10].

Since its first bibliographical appearance [11] the concept of *ubiquitous music* is not limited to simple music listening but, through the integration of sensors and other devices, aims at providing mobility, social interaction and context awareness to music creation [12]. Thus in this paper the author firstly examines the link between music listening and music creation in the general framework of active listening practices and applications (Section 2). Secondly, the author considers the potentialities of large-scale interactive environments to provide spaces devoted to music creation mainly employing the properties of a user's navigation inside an area put under the range of cameras or other motion tracking sensors. In these environments the connection between physical and perceptual processes plays an important role in that the user's movements directly influence the sound output. From a musical point of view this mechanism has been defined as "active listening" by Barry Truax in 1984 [13], and later resumed by François Pachet in 1999 [14], who considered very important the possibility of listening to music in a more creative way. Active listening applications will be briefly presented and analyzed in Section 2.1. The large-scale interactive environment system employed in the examples presented in this paper is described in Section 3.1, while in Section 3.2 the author investigates the relationship between the spatial disposition of elements on the active area and the musical concepts employed for music production. Essentially three models for the use of space are identified: the bi-dimensional slider, the interactive landmarks and the grid. Application examples will complete the presentation with further considerations in Section 4.

2 Ubiquitous Music Listening

Although the popularity of small and portable electronic devices connected to streaming internet services has greatly contributed to the progress of daily music listening, probably the phenomenon of ubiquitous music listening can be traced back to the advent and the spread of the first devices for sound reproduction

(gramophones, radios, tape recorders, etc.). In 2000 a study by John Sloboda measured the pervasiveness of everyday music listening in the 44% of probability of activities accompanied by music in each two hours period [15]. Listening opportunities range from work to personal and leisure activities such as waiting, washing, walking, running, driving, eating, shopping, relaxing and so on; personal maintenance, personal travel and leisure seem to be the most eligible moments to embed musical listening in everyday activities. Despite the variously distributed attention that characterizes this kind of musical listening, emotional involvement is by no means to be excluded, ranging from dissociation to total absorption if not trancing [16]. The influence of everyday music listening in cognition, emotions and behavior and its effect as mood modulator has been widely recognized in many studies [17], [18]. Thus if everyday listening can be considered a natural, ecologically grounded activity, the concept of ubiquitous music creation is strictly depending on the availability of advanced sound and HCI technologies and on mechanisms of connection between physical and perceptual events such as active listening.

2.1 Active Listening

Music active listening is defined as the users control of what they listen by performing meaningful actions on the musical content [14]. Music digital processing offers many ways to change the listening experience in a creative way. Some of them are the mixing of multitrack recordings [19], the song selection through music information retrieval techniques [20], [21] and [22], and the music recomposition through digital filtering [23]. Particularly interesting are active listening applications that involve the employment of a large-scale responsive environment and thus bodily interaction. In the "Orchestra Explorer" the user selects the instrumental parts to listen through deictic gestures towards different floor areas [24], whereas in "Mappe per Affetti Erranti" the musical output depends on the collaborative behaviour of the dancers such as the zone they occupy and on the level of energy of their movements [25]. A musical game called "Good or Bad?" has been proposed for a two-players interaction in a large-scale responsive environment where the task is to recompose a musical piece through multi-track recording comparison. Music active listening activities necessary to win the game include the detection of musical features such as harmonic structure, rhythm and musical meter [26].

3 Large-scale responsive environments

The author experimented with large-scale responsive environments for music production with a very simple and easily portable system that does not require to wear any marker, the "Zone Tracker" application. The system is described in Section 3.1. The user's navigation in the active area is the only interaction modality allowed in this very basic set up. Consequently, the spatial organization of the floor surface emerges as the main element for driving user interaction. The spatial models employed for various applications are analyzed in Section 3.2.

3.1 The "Zone Tracker"

The "Zone Tracker" application, implemented at the University of Padova (Italy), is composed by a ceiling mounted video camera, oriented perpendicularly to the floor and by a video module [27]. A view of the whole system is depicted in Figure 1. The camera captures the users movements inside a rectangular area, whose dimensions depend from the distance camera-floor. The active area usually is 3x4m wide. The video module analyzes the input images in various steps with the aim of obtaining well shaped blobs of the silhouette of the user seen from the top. The cartesian coordinates of the blob's barycenter are then sent via OSC [28] to MAX/Msp for audio production[1]. No other sensors or devices are included in the system which has the advantage of a high portability and adaptability to nearly every environment with the only condition of the light control. Differently from the system of the "Urban Corridor" described in [29], the "Zone Tracker" application allows the detection of the movement of a user in the active area in the form of continous data and can control a whole floor surface. For this reason the spatial organization of the floor is a central point in the application design.

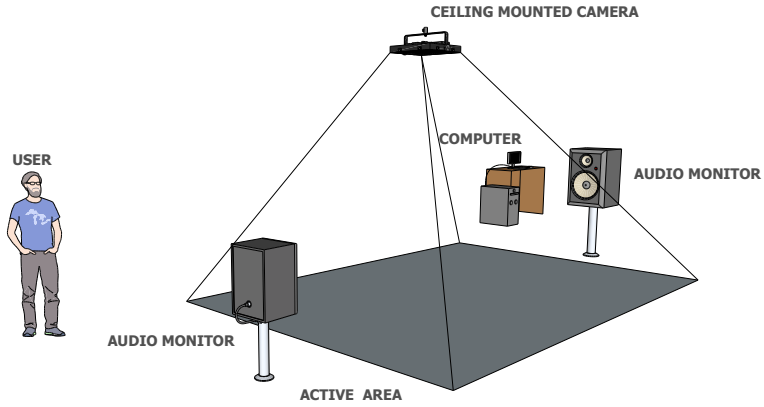


Fig. 1. A large-scale interactive environment with the "Zone Tracker" system and active area.

3.2 Spatial Models for Music Creation

Depending on the various applications and characteristics of the musical content three schemas have been identified for the use of space in the active area: the bi-dimensional slider, the interactive landmarks and the grid. The bi-dimensional

[1] `https://cycling74.com/`

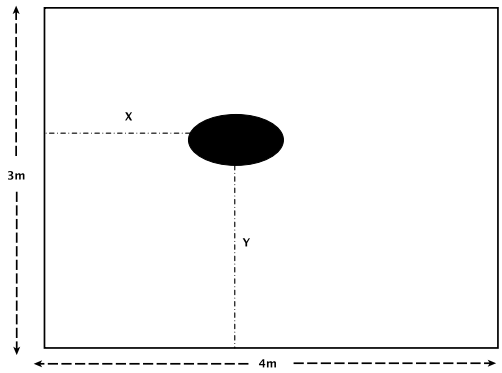


Fig. 2. Schema of a large-scale interactive environment seen from the top with a user's blob and its cartesian coordinates.

slider represents the more direct use of the user's position data, while interactive landmarks and the grid represent an ever increasing degree of constraints in the use of space. Each model will be analyzed and some example of utilization will be provided.

3.3 The Bi-dimensional Slider

The cartesian coordinates are the basic output of the system and are produced as soon as a user enters the active area (see Figure 2). The user's position may reproduce any situation where two elements represented by the x and the y values may be linked together such as the frequency and amplitude of a sinusoid or the center frequency and gain of a resonant filter. Particularly the interactive output of a filter may be used as an active listening tool for exploring the acoustical properties of a sound or musical composition. Digital filtering is also employed in auditory training therapies by Guy Bèrard [30] and Alfred Tomatis [31]. Full-body navigation in a large-scale responsive environment can then represent an added value for these therapeutic approaches, as the importance of kinaesthetic interaction and movement sonification has been widely recognized for motor and cognitive impairment rehabilitation [32]. The socialization of participants with severe disabilities is particularly important when games of imitation, mirroring and dialogue can engage more than one user and allow to share the creation experience with others [33].

3.4 The Interactive Landmarks

Interactive landmarks are bounded zones suitably delimited inside the active area. As soon as the blob of the user collides with one of these landmarks a sound is triggered or a musical process is activated. Interactive landmarks may

be freely disposed on the floor surface or may be used to represent elements with meaningful spatial features. This is the case of "Harmonic Walk", an application for the study and practice of tonal harmony for school children [34]. The harmonic space is represented by six interactive landmarks whose disposition shows the primary chords in the lower part and the parallel chords in the upper part of the schema in Fig. 3. To harmonize a tonal melody the user has to move on the interactive landmarks to trigger the musical chords that fit the melody's harmonic changes[2]. Also in this case many participatory games can be proposed by music teachers to involve the class in the activity, also while a single user is inside the active area (i.e. the class sings the song while a child accompanies it by moving on the various musical chords; the whole class mimics the movements for the melody harmonization, and so on).

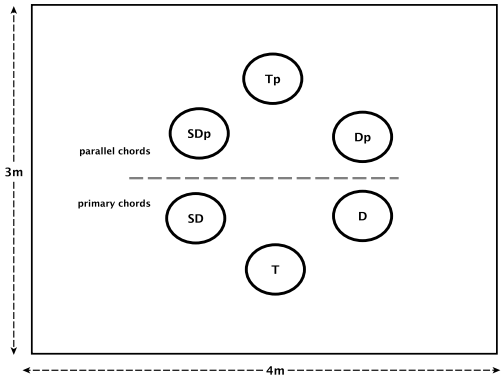


Fig. 3. Schema of the six interactive landmarks representing a tonal harmonic space in a large-scale responsive environment seen from the top with primary chords (T, tonic; D dominant; SD, subdominant) and parallel chords (Tp, parallel tonic; Dp parallel dominant; SDp, parallel subdominant).

Another example is "Interactive Soundscapes", an application that provides a virtual soundwalk by coupling ambient sounds to various interactive landmarks distributed on the active area [35]. The movements on the application's floor produce a sound output that simulates a soundwalk in a real environment[3]. Also in this case the use of a large-scale interactive environment may have important applications for music teraphy [36] and blind children rehabilitation [37].

3.5 The Grid

Many electronic music interfaces employ the framework of the grid for controlling the routing of sound for digital processing or for mixing or directly for output.

[2] `https://youtu.be/c4ru468eqM0`
[3] `https://vimeo.com/146137215`

Some of them couple also lighting effects and have been used for music creation in teraphy [38]. However, the interaction with a small physical grid is quite different from walking into a large-scale grid space. While little portable devices offer a simple way for sound routing such as pushing any button and releasing it, walking on a grid has different requirements. Firstly the user can walk only on the contiguous squares of the grid and cannot jump to those further away. Secondly if walking on a square may be the equivalent of pushing a button, what action is required to release it? The squares of the grid are physically bound to precise spatial relationships which can be used for musical purposes. An example of such constraints is "Jazz Improvisation", an application based on

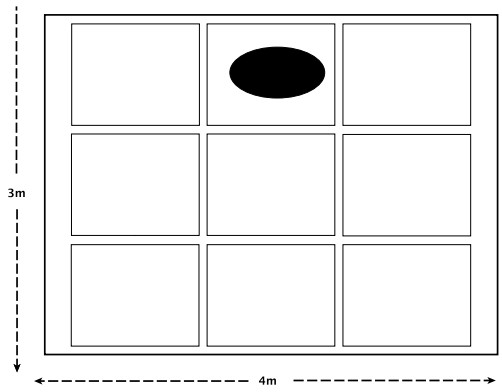


Fig. 4. Schema of a nine squares grid seen from the top with a blob occupying a single position.

a multi-track recording of a musical piece [4]. Each track corresponds to a musical instrument or musical part. The floor surface of "Jazz Improvisation" is a grid of nine interactive squares, each assigned to a single musical track. When a square is occupied for the first time, the track begins to play and remains playing in a loop until the user exits the active area; if the user goes to the same square for the second time, the corresponding track remains playing, but its volume is turned off. In case the user returns for the third time, the volume is turned on again and so on. This play/mute mechanism creates a dynamic relationship between the user's path and the composition state, which changes every time the user moves. "Jazz Improvisation" may support also a score parts listening activity, where the play/mute mechanism allows for the listening of one or more instrumental part at a time. While one child explores the score the rest of the class can guess the name of the instrument or of the group playing at the time. Another example of use of a grid is the "Resonant Memory" application for storytelling sonification [10]. Noises, music and environmental sounds are synchronized to the

[4] `https://youtu.be/uI3trfpPakU`

peripheral squares, whereas the central square is synchronized with the story. Only when the user has explored all the sounds in the peripheral squares the story begins. The use of body movement associated to spatial localization helps sound memorization and the matching of the sounds with the spoken text.

4 Conclusion

The examples presented in this short contribution represent the many possibilities and the expressivity of music making in large-scale interactive environments. These spaces have many properties which can be exploited for music creation in everyday life contexts, public events, artistic performances or therapeutic sessions. Firstly they can be easily embedded in public or private spaces. This makes them completely transparent and superimposable to the environment, reinforcing in the user the idea of naturalness and ease of use. Secondly they can make the musical creation available to bystanders subtracting it from the private sphere that traditionally characterizes both traditional and computer instruments. This new potential of sharing of the creative act must be evaluated both in its positive aspects (involvement, communication, satisfaction) and negative (fear of being judged, lack of confidentiality, insecurity, etc.). Thirdly large-scale responsive environments are really inclusive spaces. Mixing together visuospatial, bodily-kinaesthetic, musical and interpersonal intelligence [39] they can support different learning styles which can fit the profile of many users, learners or people with disabilities [40].

References

1. Keller, D., Flores, L. V., Pimenta, M. S., Capasso, A., & Tinajero, P. (2011). Convergent trends toward ubiquitous music. Journal of New Music Research, 40(3), 265-276.
2. Mandanici, M., Altieri, F., Rodà, A., & Canazza, S. (2018). Inclusive sound and music serious games in a largescale responsive environment. British Journal of Educational Technology, 49(4), 620-635
3. Mark Weiser. The computer for the 21st century. Scientific American, 265(3):94104, 1991.
4. Marianne Graves Petersen. Interactive spaces: towards a better everyday? Interactions, 12(4):4445, 2005.
5. Moreno, A., Delden, R. V., Poppe, R., Reidsma, D., & Heylen, D. (2015). Augmenting traditional playground games to enhance game experience. 2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN), pp. 140149. doi.org/10.4108/icst.intetain.2015.259399
6. Goodrich, J. & Gage, Z. (2013). Connected worlds. Retrieved February 5, 2018 from Design I/O: http:// design-io.com/projects/ConnectedWorlds/
7. Gumtau, S., Newland, P., Creed, C., & Kunath, S. (2005). MEDIATE: a responsive environment designed for children with autism. Proceedings of the 2005 International Conference on Accessible Design in the Digital World, Dundee (Australia), p. 14.

8. Birchfield, D., & Johnson-Glenberg, M. (2012). A next gen interface for embodied learning: SMALLab and the geological layer cake. In Interdisciplinary Advancements in Gaming, Simulations and Virtual Environments: Emerging Trends (pp. 51-60). IGI Global.
9. Grønbaek, K., Iversen, O. S., Kortbek, K. J., Nielsen, K. R., & Aagaard, L. (2007, June). IGameFloor: a platform for co-located collaborative games. In Proceedings of the international conference on Advances in computer entertainment technology (pp. 64-71). ACM.
10. Zanolla, S., Canazza, S., Rodà, A., Camurri, A., & Volpe, G. (2013). Entertaining listening by means of the stanza logo-motoria: an interactive multimodal environment. Entertainment Computing, 4(3), 213-220.
11. Holmquist, L. E. (2005). Ubiquitous music. Interactions, 12(4), 71-ff.
12. Pimenta, M. S., Flores, L. V., Capasso, A., Tinajero, P., & Keller, D. (2009). Ubiquitous music: concepts and metaphors. In Proceedings of the XII Brazilian Symposium on Computer Music (pp. 139-150).
13. Truax, B. (1984). Acoustic communication. Ablex Publishing Corporation.
14. Pachet, F. (1999). Active Listening: What is in the Air? Sony CSL internal Report, 1999.
15. Sloboda, J. A., O'Neill, S. A., & Ivaldi, A. (2001). Functions of music in everyday life: An exploratory study using the Experience Sampling Method. Musicae scientiae, 5(1), 9-32.
16. Herbert, R. (2016). Everyday music listening: Absorption, dissociation and trancing. Routledge.
17. Rentfrow, P. J. (2012). The role of music in everyday life: Current directions in the social psychology of music. Social and personality psychology compass, 6(5), 402-416.
18. Kassabian, A. (2013). Ubiquitous listening: Affect, attention, and distributed subjectivity. Univ. of California Press.
19. François Pachet, Olivier Delerue, and Peter Hanappe. Dynamic audio mixing. In Proceedings of ICMC, 2000
20. Masataka Goto. Smartmusickiosk: Music listening station with chorus-search function. In Proceedings of the 16th annual ACMsymposium on User interface software and technology, pages 3140. ACM, 2003.
21. Misako Goto. Active music listening interfaces based on signal processing. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 4, pages IV1441. IEEE, 2007.
22. Goto, M., Yoshii, K., Fujihara, H., Mauch, M., & Nakano, T. (2011, October). Songle: A Web Service for Active Music Listening Improved by User Contributions. In ISMIR (pp. 311-316).
23. Giovanna Varni, Gaël Dubus, Sami Oksanen, Gualtiero Volpe, Marco Fabiani, Roberto Bresin, Jari Kleimola, Vesa Välimäki, and Antonio Camurri. Interactive sonification of synchronisation of motoric behaviour in social active listening to music with mobile devices. Journal on Multimodal User Interfaces, 5(3-4):157173, 2012.
24. Antonio Camurri, Corrado Canepa, and Gualtiero Volpe. Active listening to a virtual orchestra through an expressive gestural interface: The orchestra explorer. In Proceedings of the 7th international conference on new interfaces for musical expression, pages 5661. ACM, 2007
25. Antonio Camurri, Corrado Canepa, Paolo Coletta, Barbara Mazzarino, and Gualtiero Volpe. Mappe per affetti erranti: a multimodal system for social active

listening and expressive performance. In Proceedings of the 8th Intl. Conference on New Interfaces for Musical Expression (NIME08), pages 134139, 2008.
26. Mandanici, M., Altieri, F., Pretto, N., Munaro, M., Canazza, S., & Menegatti, E. (2018, November). The "Good or Bad?" Game: Stimulating Listening Skills through Playful Engagement. In Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good (pp. 177-182). ACM.
27. Leonardo Amico. La Stanza Logo-Motoria. Un ambiente multimodale interattivo per l'insegnamento a bambini in situazione di multidisabilità. Master thesis, Dipartimento di Ingegneria dell'Informazione, Università di Padova, 2012.
28. Wright, M. (2005). Open Sound Control: an enabling technology for musical networking. Organised Sound, 10(3), 193-200.
29. Keller, D., Capasso, A., & Wilson, S. R. (2002). Urban Corridor: accumulation and interaction as form-bearing processes. In Proceedings of the International Computer Music Conference (ICMC 2002) (pp. 295-298).
30. Brockett, S. S., Lawton-Shirley, N. K., & Kimball, J. G. (2014). Berard auditory integration training: Behavior changes related to sensory modulation. Autism Insights, 2014(6), 1-10.
31. Gilmor, T. (1999). The efficacy of the Tomatis method for children with learning and communication disorders: A meta-analysis. International Journal of Listening, 13(1), 12-23.
32. Ghisio, S., Coletta, P., Piana, S., Alborno, P., Volpe, G., Camurri, A., et al. (2015). An open platform for full body interactive sonification exergames. 2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN), pp. 168175. IEEE. doi.org/10.4108/icst.intetain.2015.259584
33. Bergsland, A., & Wechsler, R. (2016). Turning movement into music: issues and applications of the MotionComposer, a therapeutic device for persons with different abilities. SoundEffects An Interdisciplinary Journal of Sound and Sound Experience, 6, 2347.
34. Mandanici, M., Rodà, A., & Canazza, S. (2016). The Harmonic Walk: An interactive physical environment to learn tonal melody accompaniment. Advances in Multimedia, 2016, 2.
35. Lionello, M., Mandanici, M., Canazza, S., & Micheloni, E. (2017). Interactive Soundscapes: Developing a Physical Space Augmented through Dynamic Sound Rendering and Granular Synthesis. In Proceedings of the 14th Sound and Music Computing Conference.
36. Viega, M. (2014, June). Listening in the ambient mode: Implications for music therapy practice and theory. In Voices: A World Forum for Music Therapy, 14(2).
37. Mandanici, M., Rodà, A. Large-Scale Interactive Environments for Mobility Training and Experience Sharing of Blind Children (2019). S. Paiva (ed.), Technological Trends in Improved Mobility of the Visually Impaired, EAI/Springer Innovations in Communication and Computing.
38. Clements-Cortes, A. (2014). Getting your groove on with the Tenori-on. Journal of Music, Technology & Education, 7(1), 59-74.
39. Gardner, H. (1983). Frames of mind: the theory of multiple intelligences. Basic books. doi.org/10.1002/ pam.4050030422
40. Mandanici, M., Altieri, F., Rodà, A., & Canazza, S. (2018). Inclusive sound and music serious games in a largescale responsive environment. British Journal of Educational Technology, 49(4), 620-635.

Ubiquitous Music, Gelassenheit and the Metaphysics of Presence: Hijacking the Live Score Piece *Ntrallazzu 4*

Marcello Messina[13] and Luzilei Aliel[2]

[1] Federal University of Paraíba, João Pessoa PB, 58051-900, Brazil
[2] University of São Paulo, São Paulo SP, 05508-020, Brazil
[3] Amazon Center for Music Research (NAP), Rio Branco AC, Brazil

Abstract. Originally composed by Marcello Messina, *Ntrallazzu* is a cycle of pieces for live score and electronics built on Max, and involving various instrumental line-ups. In particular, *Ntrallazzu 4* was performed by Luzilei Aliel on the *pífano* and electric guitar in São João del Rei during the VIII UbiMus workshop. Aliel's particular setup also involved a further layer of processing: namely, the usage of Pure Data alongside Ableton Live in order to literally hijack the original piece and open a whole set of unforeseen possibilities that abundantly transcend the original intentions. In this paper, we signify our experience by means of the concept of comprovisation, while we situate *Ntrallazzu 4* within the domain of ubiquitous music. Furthermore, we make use of the Heideggerian concept of Gelassenheit and of the Derridean concept of Metaphysics of Presence (as reformulated by Joseph Pugliese) in order to make sense of the piece.

Keywords: Comprovisation, Live score, Gelassenheit, Metaphysics of Presence.

1 Introduction

The research in ubiquitous music (ubimus) provides theoretical and methodological alternatives to proposals focused exclusively on the concepts and technological adaptations of acoustic instruments. It is important to highlight important applications in the educational field, including activities aimed at formal education (Keller, Lima, 2018, Lima et al., 2018) and the development of support strategies for musical activities in informal spaces (Ferreira et al., 2015; Keller, Lima 2016). The results of ubimus research indicate ways to overcome the obstacles in knowledge transfer in the context of activities that involve participants devoid of musical training. Another approach that has received renewed attention in ubimus research is the use and implementation of technological infrastructure outside the traditional spaces for musical making (Pimenta et al., 2012, Schiavoni et al., 2018). Among the new applications of this strategy, we can mention the works that use DIY methods to develop control mechanisms and audio processing that previously were only accessible in studio (Lazzarini et al., 2015). Recently, there have also been advances in the incorporation of the Internet of Things into musical activities (Keller; Lazzarini, 2017). Finally, a ubimus approach that can contribute to artistic achievements involves the implementation of concepts and methods based on the perspectives of

ecological cognition (Gibson, 1979; Hutchins, 2010;). This perspective encompasses the production of works involving the active participation of the audience (Basanta, 2010; Keller;Capasso, 2006), the creative use of local resources through technological support (Burtner, 2005; Gomes et al., 2014) and the use of instrumental sound sources (Aliel et al., 2015, Connors, 2015, Nance, 2007). However, there is a field of application - at the border between improvisatory practices and methods based on ecology - that still presents conceptual and procedural challenges. This field has recently been defined as the evidence-based practice linked to ecological cognition (Aliel, 2017; Aliel et al., 2015).

In this work we will focus on this last aspect of ubimus research, that seeks to transcend the rigid separation of roles and social practices in which (Western) music can be practiced and understood, considering the use of technological devices that have the potential to guarantee universal access to the production and consumption of music. This problem is linked to the acoustic-instrumental paradigm, that will be discussed in more details in the next subsection. In this context, Keller et al. illustrate their perspective on ubiquitous music:

> Previous musical practices provided the safe refuge of instruments as the physical support for all sound producing actions. These actions could be encoded as a series of discrete symbols - a score - which would guide the performers through a finite set of possible interactions with their instruments. Performances would occur within a space especially designed for musical activities - the concert hall - guaranteeing acoustic characteristics compatible with instrumental sound source power and projection. Furthermore, a crisp separation between performers and public, following an established ritualized set of actions - play / listen, bow / applaud - reinforced by the physical separation between stage and audience seats, allowed for strictly predefined roles in music making: musicians play, spectators just listen. Most of this social paraphernalia breaks down in the context of ubiquitous musical practices (Keller et al. 2010, p. 320).

In *Ntrallazzu 4* we try to relate compositional structures with the adaptive processes made via improvisation, and to conceptualize how this type of ubiquitous artistic practice can help to alleviate the impact of this segregative "ritual" associated to the artist/audience model and the acoustic-instrumental paradigm. However, it is necessary to understand how the acoustic-instrumental paradigm occurs.

1.1 **Acoustic-Instrumental Paradigm**

Etymologically, the term “paradigm” originates from the Greek *paradeigma* which means a model or pattern, corresponding to something that will serve as an example to be followed in various situations. The social norms that regulate the behaviour of any human group set precise limits and determine how each individual should act within those limits. Often, paradigms are established as dogmas that can be transmitted for political reasons, or that in some cases are used in human interactions to increase social cohesion. In the specific case of the acoustic-instrumental paradigm - cf. critical discussions in Bown et al. (2009), Keller (2000), Keller (2014) and Lima et al. (2018) -, this is a normative and substantially Eurocentric concept that has a dramatic impact on creative musical practices. In this way, creative agendas focused

on the objectives of instrumental practice, relegating to a second level the cultural manifestations that were perceived as being external to this type of practice (cf. critical discussions in Bown et al., 2009; Keller, 2014). This conceptualization induces, at least in a large part of individuals, an understanding of music making as being limited to a few talented and formally trained individuals. According to Wishart (2009), the use of technology has become essential to the pursuit of creative products based purely on sound, fostering a view of music focused on the acousmatic phenomenon. However, the construction of tools centered on instrumental models tends to reduce interaction strategies based on the exploration of the potential of sound, limiting the possibilities of action to interfaces that emulate acoustic instruments. It also limits the use of local material resources in aesthetic decisions, as it imposes the production of contents that may be not necessarily related to local social, cultural and histotical contexts. Emphasis is placed on the software, system and machine embodied in the instrument. The digital musical instrument thus becomes a new fetish. Technological resources serve as accessories for old acoustic-instrumental practices focused exclusively on the aural properties of the instruments.

Taking this problem as a starting point, we propose the critical use of Heidegger's concept of Gelassenheit, as adapted to sound practices by Aliel et al (2018) in order to construct a significant speculative basis to understand the proposed processes in *Ntrallazzu 4*, and the ways in which we can transcend the paradigms mentioned above.

1.2 **Gelassenheit**

In an attempt to explore alternatives to the acoustic-instrumental paradigm, we will seek a theorization of the philosophical concept of Gelassenheit for the field of musical practices. Gelassenheit is a term coined by Heidegger (1966). Its literal translation would be something like "serenity", but Heidegger's formulation transcends the literal meaning of the word. What Heidegger proposes is that Gelassenheit is a stage to be achieved through an openness to new forms of thought. In this wake, the author proposes two thought-forms: 1. Calculating thinking, which is understood within a "scientific-artistic method" with the purpose of measuring, collecting data and reproducing results. According to Heidegger, the use of new technologies is centered on calculating thinking. 2. Meditative thinking, which is the aptitude to be open to unpredictable actions, to unexpected events, and to mystery itself (Heidegger, 1966). From an artistic perspective, Aliel (2018) considers this last behavior as a process of adaptation and modification within self-reflective strategies (Donald, 2006). In this way, meditative thinking is not uniquely bound up with the product (as in calculating thought-form), but it is focused on experience in a particular way.

It is with this kind of unique experience that disparities - both in terms of technical means and of consolidated knowledge - are reduced, allowing a greater socialization of artistic practice. Adaptabilities generate products, but these come from the openness of agents towards moments of unpredictability and the provision of

reactions to these new contents. When we deal with this concept, that goes beyond the essence of calculative thinking, we try to find specific moments in the creative process (detouring, Keller and Lazzarini, 2017) where control is eliminated or reduced, enabling unpredictable conditions.

The etymology of the term paradox is based on the Greek *paradoxon*, also found in the late Latin word *paradoxum*. The word is composed of the prefix *para-*, which means "contrary to," "altered," or "opposed to," in conjunction with the nominal suffix *-doxa*, meaning “opinion”. In creative practices, Aliel (2018) conceptualizes paradoxes as simulacra, considered as means to understand acoustic-instrumental paradigms as a process of adaptation to new behavioral/environmental conditions, stimulating greater possibilities of material resources.

Therefore, the objective would not be increasing technical or methodological efforts in order to guarantee the exact repetition of what was planned (as in instrumental virtuosity - simulation), but rather offering significant references in a way that each agent can find their artistic singularity (simulacrum), be it coextensive or not with traditional artistic processes (Costa, 2016).

In short, as an alternative to calculative thinking, we propose to adopt Heidegger's path of meditative thinking in order to generate paradoxes that place the decision-making process outside the acoustic-instrumental paradigm (Aliel et.al 2018). The absence of control acts as impetus to arrive at unexpected results, placing the artist in an atypical frame of possibilities and forcing them to adapt their behavior to new, previously non-existent contexts. Thus, with the absence of control, the artist acquires space to introduce divergent conceptions of the expected results within their preexisting knowledge.

The reflection on calculating thought as generator of resources based on explicit knowledge and on meditative thinking as a procedural strategy that uses elements of calculating knowledge allows for the incorporation of Gelassenheiten into creative cognitive-ecological practices, coextensive with ubimus research. However, one of the problems would be: how can artists exploit material resources by fostering playful strategies for discovering significant artistic materials? The question is complex because it pushes creative practices out of the tradition of acoustic-instrumental thinking. The association between acoustic instruments and musical structures makes the experience more familiar to musician-instrumentalists. Schiavoni et al. (2018) cite a phrase by Trueman comparing the orchestra of acoustic instruments with the orchestra of laptops and arguing that "even though it is somewhat different, its goal is not at all different from a traditional orchestra in what concerns to musical ability." In addition to the literal transcription of acoustic instrument interactions, the orchestras reproduce the layout of the Italian stage, emphasizing the separation between creative participants (the musicians) and the passive audience (Princeton Laptop Orchestra or Stanford Laptop Orchestra) which reproduce the same model.

In summary, we propose the adoption of Gelassenheit within the field of creative cognitive-ecological practices aligned with the proposals of the ubimus research. This concept involves several components that can be thought of as factors linked to creative practices, encompassing cognition, materiality, social organization and the use of computational resources. If the ubimus proposals can serve to advance the

musical concepts in parallel with the technological advances, it is necessary to reconsider and alleviate the hegemony of the acoustic-instrumental paradigm. As shown by various projects based on the ecological approach, the riddance of such a paradigm does not imply the exclusion of acoustic instruments as musical tools or as sound resources (Aliel et al., 2015, Connors, 2015, Nance, 2007)

2 *Ntrallazzu*

Composed by Marcello Messina, *Ntrallazzu 4* was performed by Luzilei Aliel on the *pífano* (Brazilian traditional fife), electric guitar and audio processing (delay, reverb, granular synthesis, resonators and etc). The performance took place at the Universidade Federal de São João del Rei, in Brazil, during the VIII UbiMus - Eighth Workshop on Ubiquitous Music, on 14 September 2018 (Messina and Aliel, 2018).

As suggested by its very title, *Ntrallazzu 4* is the fourth instance of a cycle of pieces, all titled *Ntrallazzu*. In Sicilian, the word "ntrallazzu" refers to the practice of smuggling locally produced state rationed crops during World War II: this was a practice of resistance criminalised by the Italian authorities, that was aimed at contrasting indigence in Sicily. By establishing micro-patterns of interaction between themselves and with the live electronics, the performer(s) of each piece of the Ntrallazzu cycle symbolically reproduce(s) the secret exchanges that disobeyed a violent regime of state control and punishment over the lives and means of subsistence of the islanders. After WW2, the Sicilian term, literally denoting this type of illegal resistance from below, has been substantially emptied of its original meaning and Italianised in the form "intrallazzo", that at nationwide level refers to political corruption and bribes (Di Capua, 2005. p. 305).

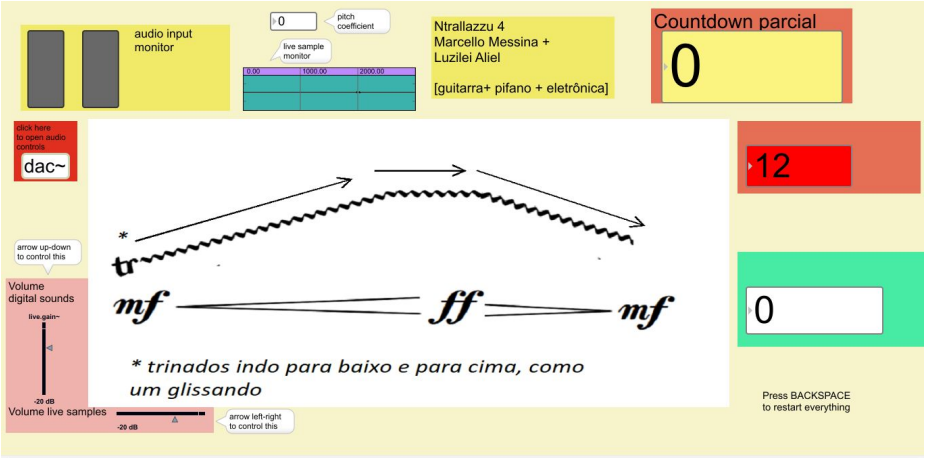


Fig. 1. A snapshot of the general GUI that is presented to the performer in *Ntrallazzu 4*

All the pieces of the *Ntrallazzu* cycle are based on a projected score that interacts in real time with the material played by the performer(s). While one of the performers plays, the sound is fed to and processed by patching software, and generates both electronic sounds and a score composed of preloaded graphical fragments, which is generally, but not necessarily, to be performed by a second player. Both the live score and the electronics run on Max (fig.1).

2.1 **Metaphysics of presence**

Intrinsic to the patch algorithm of *Ntrallazzu* is the very simple detection of instrumental sound and the discardure of background noise coming from the room and the audience. In this way, through simple operations of calibration, the patch crystallises some of the very discursive binaries that, in other ways, *Ntrallazzu* as a cycle intends to challenge and question. By calibrating the patch in order not to interact with the "accidental" sounds produced by the audience, the piece explicitly confirms the operativity of these binaries: namely, sound vs. silence and instrumental sound (wanted) vs. background noise (unwanted), hence marking the fundamental border between performers and audience.

Furthermore, the live score is triggered by sonic activity from the instrument(s), and ends once they stop playing. Now, this is a rudimentary form of biometric detection that boils down to a recognition of the "living presence" of the performers, as well as a detection of their disappearance once they stop playing. This reliance on what Joseph Pugliese, after Derrida, calls "metaphysics of presence" (Pugliese, 2014), is potentially very problematic, both theoretically and practically. How can one distinguish physical "presence" from "non-presence" if the detection is based on electronic impulses that might well simulate the two states? Is it not always already a simulation that is at stake? In this sense, how we distinguish real simulation from fake simulation? All this, in practical terms, makes Ntrallazzu always vulnerable and somewhat unstable, a characteristic that is definitely a distinctive part of the piece.

Finally, and taking another cue from Pugliese, we need to acknowledge that the bodies of the musicians have "already been technologised" (Pugliese, 2014, p. 665) before the biometric detection operated by the patch. First, that is because the human/instrument combination is a fundamental interaction that is already technological, and that, in the narratives that inscribe (Western) art music, traces a fundamental border between musician and listener. The instrument, here, awards access to some subjects while simultaneously preventing everyone else from accessing music making. More in general, the a priori "technologisation" that marks the bodies of the musicians even before the addition of live electronics might be understood as part of what Pugliese & Stryker call "somatechnics" (2009), that is, the intersection between the body as a physical, natural object and the very same body as a discursive, biocultural artifact that is always determined socially. Somatechnics makes the task of detecting physical presence through physical sound even more problematic, and confirms that Ntrallazzu, as a work of art, remains a highly fragile and unstable construction.

2.2 **The performance of *Ntrallazzu 4***

Of the five pieces that compose the cycle so far, *Ntrallazzu 1*, *Ntrallazzu 2* and *Ntrallazzu 5* call for a duo, whereas *Ntrallazzu 3* and *Ntrallazzu 4* call for a single performer. In Ntrallazzu 4, however, Aliel played both the *pífano* and the electric guitar and, thanks to a specific use of the effects, managed to maintain the feeling of an interaction between the two instruments. Initially the proposal of *Ntrallazzu 4* is to capture ambient sounds, be them instrumental/noise and/or related aural inputs, in a generic way, considering intentional and unintentional actions by various agents, such as performers, audiences and the like. This premise creates a condition of sound reception and transformation into notation (as explained in section 2.1). The performer/comproviser, however, added another layer of information feedback, creating a secondary path to the resulting sound, in a way to "hack" the first system.

For a field of comprovisation, used in the performance approach of *Ntrallazzu 4*, it was necessary to define guideline plans and contingency plans (Aliel, 2017). Guideline plans are rules or definitions that are not likely to be modified during the performance: that is, from a musical perspective, the composition; or, from a computational perspective, the algorithm. In *Ntrallazzu 4*, a second technical feedback system was used to create more complexity and allow for unexpected elements (*Gelassenheit*), as observable in some artistic works stemming from discussions on grounded creative practices (Aliel, 2018). Aliel curated the creation and organization of the original patch and introduced a new patch on Pure Data (PD) that connects via *jackaudio* to Ableton live software. In this way, two microphones are positioned next to the speakers that reproduce the original signal of the Ntrallazzu 4 patch; this signal is then transductively transformed into MIDI protocol language, via the [ftom] - (frequency for MIDI) object in PD. Obviously, the resulting algorithm is not limited solely to this object, however it constitutes its fundamental axis: its structure is based on signal captured via pitch detection, similar to the procedures found in digital tuners. With the numeric MIDI data, the transformation is subdivided into four channels controlled by a stochastic algorithm that addresses each channel for sound processing in Ableton Live. That is, during the performance there is no knowledge about which sonic processes will be applied, since the entire process is randomly controlled by the machine, introducing aspects consistent with the *Gelassenheit* strategy (Aliel, 2018).

In the guideline plan, Aliel considered the organization of the original *Ntrallazzu 4* patch and introduced a new Pure Data (PD) patch that connects via *jackaudio* with Ableton live software. Briefly, the PD patch captures the sounds reproduced in the amplifiers of the first *Ntrallazzu 4* patch and transforms the signal into MIDI protocol. This transformation is subdivided into four channels that are controlled by a stochastic algorithm that addresses each channel for processing in Ableton Live. That is, during the performance there is no knowledge about what types of processing, whether involving pitch or types of dynamics will be reproduced, as the whole process is randomly controlled by the machine.

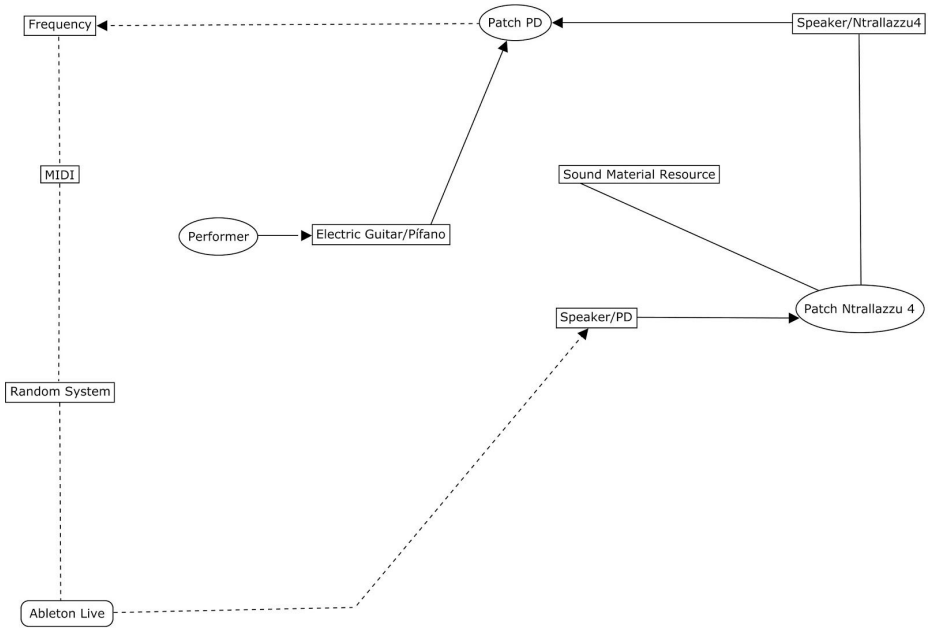


Fig. 2. Flowchart of the *Ntrallazzu 4* comprovisation system

On the other hand, the contingency plans of *Ntrallazzu 4* are limited to the process of "hacking" and adaptation to the constructed systems. Contingency plans in compositions refer to elements that are not previously organized and must be solved at the moment of performance: in musical structures this is associated with improvisation, while in computational structures it is associated to complex dynamic systems, or stochastic resources. In *Ntrallazzu 4,* the adaptation process consists in dealing with the overlapping of feedback layers. At each new sound layer, two distributions are conducted separately but provide parallel effects. While a patch uses the data captured to generate musical notation content (score) the other patch uses the transduced information to select the audio processing. The contingency process therefore involves the adaptation of the performance to notation and loudness events, and must be "solved" in real time. It is important to point out that with each new adaptation of the performer, a new layer of sound information is introduced to the system, allowing for a change with a greater or lesser impact on the final result. Ultimately, the machine participates in the work effectively. In the contingency plan, Aliel handles multiple textural layers, some of which are more easily recognizable, while others become fully diffused. This lack of standardization requires a high level of openness to events by the performer, and the acceptance that the final material may be very different from that imagined by the composer.

3 Final Remarks

In all the pieces of the *Ntrallazzu* cycle, the "biometric detection" of the presence and absence of the performer(s) is always at risk of either being disrupted by unpredictable circumstances, such as audience noise, etc., or of being deliberately simulated or dissimulated by the musicians.

In *Ntrallazzu 4*, however, disruption is taken to a further level, in that the original piece, with its set of functional and aural predictions, is literally hijacked by the performer/comproviser and fed to a whole set of new processings, therefore opening a wide range of unforeseen possibilities. This resonates again with ubiquitous music and with the transcendence of the rigid separation of roles (in particular, composer vs. performer) and social practices on which (Western) art music is predicated, through the use of technological devices that have the potential to grant universal access to the making and consumption of music.

Situating *Ntrallazzu 4* and the whole *Ntrallazzu* cycle as ubiquitous music also involves assessing the ways in which they blur the separation between performers and audience by projecting the score and letting spectators appreciate the interaction. In this way, the "mysterious" ritual of score reading, that is normally negotiated among the performers, is opened to the general public. Importantly, this is not "any" score reading, as the live score involves a high degree of unpredictability that may engage the performer(s) and audience in a sort of "interpassive" interaction (Reuben, 2014). In turn, this allows for a mitigation of the strict protocol of ceremonial actions such as bowing and applauding, as the score flags the end of the piece, leaving no room for surprise awkward doubts.

Ironically, of all the various circumstances in which pieces of the *Ntrallazzu* cycle were performed, it was precisely on the occasion of the performance of *Ntrallazzu 4,* reiteratively, during the VIII UbiMus workshop, that this strict protocol was implicitly reestablished by the coercive – if bona fide – action of a sound/video technician. Once he saw fragments of a score on the screen, the technician decided to turn off the projection in the middle of the performance. Supposedly, even in that context, the "mystery" of score reading was deemed to belong too exclusively to the performer, in a way that any form of audience participation in the process was unmistakably categorized as an error.

Just to clarify, disruptions such as this one are to be intended as an integral part of the "experimental" rubric that inscribes this piece as well as its technical preparation, compositional process and performance-oriented training. In this sense, the fact that the projection was arbitrarily interrupted does not represent a sort of "failure" of the piece – on the contrary, it proves one of the fundamental points made by the piece itself. Furthermore, if on the one hand the interruption of the projection seems to undo precisely the philosophical premises on which ubiquitous music is predicated, on the other hand, as the deliberate intervention of an agent that exists outside the composer/performer dyad, it may also be categorised as a form of "breaking down" of

the "social paraphernalia" that characterize traditional musical practices. In general, unexpected disruptions are a fundamental part of the whole *Ntrallazzu* cycle.

References

1. Aliel, L. Ensaios sobre comprovisações em ecologia sonora: Perspectivas práticas e teóricas. Dissertação de Mestrado em Música. São Paulo: USP, 2017.
2. Aliel, L.; Keller, D. & Costa, R.: Comprovisation: An approach from aesthetic heuristics in ecocomposition (Comprovisação: Abordagens desde a heurística estética em ecocomposição). In: *Proceedings of the Brazilian Symposium on Computer Music (SBCM 2015)*, 169-180. Campinas, SP: SBC (2015).
3. Aliel, L.; Keller, D. & Costa, R.: Theoretical perspectives for the analysis of ecologically grounded creative practice (Perspectivas teóricas para a análise das práticas criativas ecocognitivas). In: Damián Keller, Helena Lima (eds.), *Ubiquitous Music Applications* (*Aplicações em Música Ubíqua)*. São Paulo, SP: ANPPOM (2018).
4. Basanta, A. Syntax as sign: The use of ecological models within a semiotic approach to electroacoustic composition. Organised Sound 15 (2), 125-132, 2010. (Doi:10.1017/S1355771810000117.).
5. Bown, O.; Eldridge, A.; Mccormack, J. Understanding interaction in contemporary digital music: From instruments to behavioural objects. Organised Sound 14, 188-196, 2009. (Doi: 10.1017/S1355771809000296.).
6. Connors, T. M. Audiovisual Installation as Ecological Perfomativity. In: Proceedings of 21st International Symposium on Electronic Art (ISEA 2015). Vancouver, Canada: ISEA, 2015.
7. Costa, V. F. Morfologia da Obra Aberta: esboço de uma teoria geral da forma musical. Editora Prismas, Curitiba, 2016
8. Di Capua, G. Il biennio cruciale (luglio 1943-giugno 1945): l'Italia di Charles Poletti. Rubbettino Editore, Soveria Mannelli, (2005)
9. Donald, M. Art and Cognitive Evolution. In M. Turner (ed.), The Artful Mind. Oxford: Oxford University Press, (2006)
10. Ferreira, E.; Keller, D.; Farias, F. M.; Pinheiro da Silva, F.; Lazzarini, V.; Pimenta, M. S.; Lima, M. H.; Costalonga, L. L.; Johann, M. Marcação temporal em ambientes domésticos e comerciais: Estudo comparativo. In: D. Keller; M. A. Scarpellini (eds.), Anais do Simpósio Internacional de Música na Amazônia (SIMA 2014), Vol. 2. Rio Branco, AC: Editora da UFAC, (2014).
11. Gibson, J. J.: *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin (1979).
12. Heidegger, M. Gelassenheit. New York, NY: Harper Collins, 1966.
13. Hutchins, E.: Cognitive ecology. *Topics in Cognitive Science* 2 (4), 705-715 (2010).
14. Keller, D. Compositional processes from an ecological perspective. Leonardo Music Journal 10, 55-60,. (Doi: 10.1162/096112100570459).(2000)
15. Keller, D. Characterizing resources in ubimus research: Volatility and rivalry. In: Proceedings of the V Workshop in Ubiquitous Music (V UbiMus). Vitória, ES: Ubiquitous Music Group, 2014.
16. Keller, D., Barreiro, D. L., Queiroz, M., Pimenta, M. S. Anchoring in ubiquitous musical activities, in Proceedings of the International Computer Music Conference, University of Michigan Library, Ann Arbor, 319-326 (2010).

17. Keller, D.; Capasso, A. New concepts and techniques in eco-composition. Organised Sound 11 (1), 55-62, 2006. (Doi: 10.1017/S1355771806000082.)
18. Keller, D.; Lima, M. H.; Práticas cognitivo-ecológicas em Ubimus: Sons do CAp. In: Anais do Workshop em Música Ubíqua (UbiMus 2018). São João del Rei, MG: UFSJ. (2018).
19. Keller, D. & Lazzarini, V. Ecologically grounded creative practices in ubiquitous music. Organised Sound 22 (1): 61–72, (2017).
20. Keller, D.; Lima, M. H. Supporting everyday creativity in ubiquitous music making. In: Trends in Music Information Seeking, Behavior, and Retrieval for Creative (pp. 78-99). Vancouver, BC: IGI Global, (2016).
21. Lazzarini, V.; Keller, D.; Kuhn, C.; Pimenta, M.; Timoney, J. Prototyping of ubiquitous music ecosystems. Journal of Cases on Information Technology (17), 73-85, (2015). (Doi: 10.4018/JCIT.2015100105.)
22. Lima, M. H.; Keller, D.; Pimenta, M. S.; Lazzarini, V.; Miletto, E. M. Creativity-centred design for ubiquitous musical activities: Two case studies. Journal of Music, Technology and Education 5 (2), 195-222, (2012). (Doi: 10.1386/jmte.5.2.195_1.) (2018)
23. Messina, M.; Aliel, L. Ntrallazzu 4 [a cycle of pieces for extractable parts, live scores and electronics]. Workshop em Música Ubíqua / Ubiquitous Music Workshop (UbiMus 2018), São João del Rei, Brazil, (2018).
24. Nance, R. W. Compositional explorations of plastic sound. Doctoral Thesis in Music, De Montfort University, UK, 2007.
25. Pimenta, M. S.; Miletto, E. M.; Keller, D.; Flores, L. V. Technological support for online communities focusing on music creation: Adopting collaboration, flexibility and multiculturality from Brazilian creativity styles. In N. A. Azab (ed.), Cases on Web 2.0 in Developing Countries: Studies on Implementation, Application and Use. Vancouver, BC: IGI Global Press, (2012). (ISBN: 1466625155.)
26. Pugliese, J. The alleged liveness of "Live": Legal visuality, biometric liveness testing and the metaphysics of presence. In: Wagner, A., Sherwin. R. K. (eds.) Law, culture and visual studies, pp. 649-669. Springer, Dordrecht (2014).
27. Pugliese, J., Stryker, S. The somatechnics of race and whiteness. Social Semiotics, 19(1), 1-8. (2009).
28. Reuben, F. Reworking Musical Strategies in the Digital Age. PhD Thesis: University of York (2011).
29. Schiavoni, F. L.; Silva, E. X.; Cançado, P. G. N.. Orchidea: Uma orquestra de dispositivos móveis. In: Anais do Workshop em Música Ubíqua (UbiMus 2018). São João del Rei, MG: UFSJ. (2018)
30. Wishhart, T. Computer music: Some reflections. In R. T. Dean (ed.), The Oxford Handbook of Computer Music, 151–160. New York: Oxford University Press, 2009.

Live Patching and Remote Interaction: A Practice-Based, Intercontinental Approach to Kiwi

Marcello Messina[134] , João Svidzinski[2], Deivid de Menezes Bezerra[3], David Ferreira da Costa[3 4].

[1] Federal University of Paraíba, João Pessoa PB, 58051-900, Brazil
[2] University Paris 8 - Musidanse/CICM, 93526 Saint-Denis, France.
[3] Federal University of Acre, Rio Branco AC, 69920-900, Brazil
[4] Amazon Center for Music Research (NAP), Rio Branco AC, Brazil

Abstract. This paper introduces, documents and reflects on an intercontinental live patching experience based on simultaneous remote interaction using the software Kiwi, and that can be subsumed under several features of Ubiquitous Music. The experience involved two academic groups based in three different universities between Brazil and France, namely, a research group from the two Brazilian Federal Universities of Acre and Paraíba, and a working group based at the University Paris 8 in France. The intercontinental simultaneous interaction may trigger reflections on the implications of the presence/absence of the human being, on the implicit patterns of territorialisation reproduced in the context of intercontinental live patching, and on the operative action of mnemonic processes within the practice.

Keywords: Music composition, epistemology of music composition, computer music, Live Patching, Remote Interaction, Kiwi.

1 Introduction

This paper introduces, documents and reflects on an intercontinental live patching experience based on simultaneous remote interaction using the software Kiwi. Previous diffused live patching experiences include a proxy for the Pure Data community, called *peer data*, that was developed by IOhannes m zmölnig, and allows multiple users to concurrently intervene on the same patch; *peer data* was successfully used in the context of the *blind date* project by Pd~Graz, which will be discussed further on. There are other approaches to remote collaboration on Pure Data, such as the project *Destino Pirilampo*, by Luzilei Aliel da Silva and José Eduardo Fornari Novo Junior [1], where remote users send signals to a unified central patch, but do not have control on the patch itself.

This experience can be subsumed under several features of Ubiquitous Music. First of all, according to the understanding suggested by Luzilei Aliel and José Fornari, it qualifies as a particular type of "electroacoustic music, in which electronic devices are now ubiquitously interconnected to make music together" [1]; secondly, as Keller *et al* maintain, this very same "multiplicity of interconnected devices" replaces the operativity of "identifiable musical instruments" as well as the complex "social paraphernalia" (scores, concerts, set rituals, a clearly defined audience separated by the musicians, etc.) that characterise typical musical practices [2]; finally, the particular set

of material circumstances involved in the experience implies the encounter and clash between the human presence/absence and a characteristic territorialisation of creative actions, that evokes Keller and Lazzarini's formulation of "ecologically grounded creative practice" [3].

The Kiwi project originated from a social, scientific and musical context that is in constant development. Interaction and collaboration through digital media are taken as key concepts in the current social context. Cloud computing permits to explore the potential of digital processing and storage of information through the mediation of communication networks – mainly the internet.

The project ANR MUSICOLL[4] (2016-2018), in partnership with the CICM [5] and the private company Ohm Force [6], conducted research centered on collaborative and portable real-time music, with the development of a new tool for this practice: the software Kiwi [7]. After the finalisation of the project in 2018, all these objectives were successfully achieved. However, the creative potential of Kiwi was scarcely considered.

The practice of live patching functioned as a starting point for a new approach to Kiwi. The principle of starting a patch "from scratch" in a collaborative, simultaneous practice, working together on the same patching canvas from different parts of the world will hopefully help the visualisation of a series of original properties belonging to this new software. This would eventually let a series of new practices emerge, freely from the canons and prejudices of dominant academic schools of composition.

This project was initiated by the association of two different groups, that collaborate with the same tools of digital musical creation: namely, the research project Live/Acc/Patch, based in two different Brazilian Federal Universities (Acre and Paraíba), and the working group associated to the module *Introduction à la programmation avec Kiwi, Max et Pure Data 1*, from the University Paris 8.

This collaboration, therefore, unites two different countries, France and Brazil, as well as two different conceptions of music-making, with their own intrinsic know-hows. The use of the software Kiwi as a shared tool constituted the intersectional space of this collaboration, that resulted in a particular medium-specific practice of music-making, distinctive of this interinstitutional, border-crossing app.

2 Live/Acc/Patch

Live/Acc/Patch is a research project initiated in 2018 through the mutual finding of the Brazilian National Council for Scientific and Technological Development (CNPq) and the Federal University of Acre (UFAC), as part of the national Institutional Programme for Undergraduate Research (PIBIC).

The project aims to explore viable connections between collaborative live patching and the learning of patching programs such as Pure Data [8]. A first part of the project involved practising collaborative live patching with users already initiated with Pure Data. A second, ongoing part of the project consists in working closely with users not initiated to patching software, with the aim of transferring to them basic notions of live patching.

2.1 Live coding and live patching

The term "live patching" derives from the concept of "live coding", defined by Thor Magnusson as a "form of musical performance that involves the real-time composition of music by means of writing code" [9]. Live coding emerged as a creative practice and as a research area in music in the last 15-20 years. Collins *et al* situate live coding within a new performance area based on risk, unpredictability and, therefore, in the possibility of unexpected developments [10]. Magnusson illustrates the transition from score to algorithm heralded by live coding as a new form of creation of musical meaning [9].

While live coding is a general term that may (or may not) evoke TUIs, or textual user interfaces, live patching refers to the same practice applied to GUIs, or graphical user interfaces [11]. The increased appeal and smoother learning curve granted by GUIs makes them more indicated for the collaboration and educational purposes of the Live/Acc/Patch project. In terms of previous collective live patching experiences, a crucial reference is represented by the *blind date* and *rec.wie.m* projects by the Pd~Graz association [12][13]. Even with this advantage, the project demands the formulation of strategies aimed at facilitating the understanding of patching languages and accelerate learning for users without any experience of this type of apps.

3 From Pure Data to Kiwi

Soon after its start, the Live/Acc/Patch project partially relocated to the Federal University of Paraíba, in João Pessoa, some 4,000 km away from Rio Branco. There, a collaboration with the University Paris 8 was soon established, based on the use of Kiwi. So far, Pure Data had been used as the only patching program for the project. This involved various challenges, involving mainly the impossibility (or difficulty) of working simultaneously on a patch. Initially, then, the group had started working collaboratively on live patching practices, interacting consecutively in turns on the same patch. Later, after splitting between the states of Acre and Paraíba, group members had started working remotely, each of them saving patches on shared cloud folders and intervening on structures already initiated by the others.

The collaboration with Paris 8 introduced the possibility of using Kiwi as a tool that could effectively alleviate the difficulties connected to the huge distances that inscribed collaborative interaction in the project. Project members soon decided that Kiwi would by no means supplant Pure Data, and that the two, on the contrary, would coexist within the project.

3.1 Characteristics of Kiwi

The collaboration model offered by Kiwi is centred on the sharing of the working space, that is, the patch canvas; however, the control parameters and the audio signals are kept only locally. The extreme similarities between Kiwi and Pure Data helped Live/Acc/Patch members to painlessly transition from one language to the other and back. An important characteristic of Kiwi consists in the presence of rooms where collaborative patches can be created and put at the disposal of the other members of the network. All users connected to the server have access to a list of rooms: they may download, delete, upload, rename, duplicate patches from this window. Most

importantly, each room allows for multiple users to be connected at the same time, each making real-time modification on the patch. In fact, the instances of intercontinental collaboration discussed in this paper happened, each of the times, within the limits of a specific room.

4 The Module *Introduction à la programmation avec Kiwi, Max et Pure Data 1* at University Paris 8

The second working group that participated in the international collaboration was the class of the module *Introduction à la programmation avec Kiwi, Max et Pure Data 1*, offered to *Licence 2* (second year undergraduate) students of the music department with CAO (computer-assisted composition) specialisation, of the University Paris 8. The module teaches basic notions of modular graphic syntax for audio synthesis and processing, that is, the basic principles of Max, Pure Data and Kiwi. The course also covers notions of digital signal processing and electroacoustic composition.

4.1 Learning with Kiwi at the University Paris 8

The 2019 class that participated in the international collaboration is composed of roughly ten students, the majority of whom is a total beginner in terms of computer-assisted composition. The first eleven lessons covered notions of digital processing and an introduction to electroacoustic composition, including frequency and amplitude modulation, delay, flanging, pseudo-flanging and sampling. Kiwi was only introduced during the seventh lesson, when students had already familiarised with basic principles of patching. For them, the specific aim of the international collaboration consisted in the opportunity of an additional creative activity, useful to complement classroom learning.

In addition, two patches were created collaboratively by the class, independently from the international collaboration. These two patches focussed on frequency modulation, and are available on the Kiwi server as KiwiMaxPd2019 and KiwiMaxPd20192. This last patch contains four instances of a basic model of John Chowning's FM synthesis [14].

5 Simultaneous Intercontinental Live Patching

After a series of local practices intended to familiarise quickly with Kiwi as a patching environment, towards the end of 2018, the members of the Live/Acc/Patch project started live patching between the two Brazilian states of Acre and Paraíba. In April 2019, three attempts at intercontinental live patching between France and Brazil were made.

5.1 The first interstate sessions in Brazil

A first set of interstate sessions occurred between Acre and Paraíba in the last months of 2018. Sessions are available on the Kiwi server as LIVE-ACC-PATCH

UFAC/UFPB followed by consecutive numbers. A total of four sessions was created between October 17th and November 28th.

5.2 Intercontinental appointments

Relevant intercontinental session included a test session on April 11th (available as LiveKiwi 11-04-2019), three smaller group rehearsals on April 17th (available as kiwitest 17.04.19; kiwitest_2 17.04.19; kiwitest 17.04.19 (versao 2)), and a final collaborative event on April 18th (fig 3), which was split into consecutive sessions (available as LiveKiwi 18-04-2019 and LiveKiwi_v2 18-04-2019). Importantly, due to some repeated server failures during the test session on April 11th, participants based in Brazil resolved opening a separate session room on that day (available as ACRE-PARAÍBA TEST) that eventually worked smoothly. Thanks to the help of the Kiwi developers, the server problem did not reappear on the following session days.

5.3 Reflections on Kiwi, live coding and simultaneous remote interaction

The different sessions mentioned above were useful in order to formulate critical and analytical reflections on the set of operativities, functionalities and conceptual nodes that are at stake when opting for using Kiwi.

The practical and conceptual limits of simultaneous remote interaction. The most appealing feature offered by Kiwi is, by far, the idea of collective interaction on patching being potentially simultaneous and remote [15]. While we have abundantly enjoyed these two characteristics along the various sessions, the aforementioned problems that occurred upon the first intercontinental test on April 11th triggered a broader reflection on the unconditional desirability of simultaneity in the context of remote interaction. In other words, is it always absolutely necessary to interact right at the same time? More importantly, is it always safe and reasonable to rely on simultaneity?

Kiwi does in fact allow for simultaneity to be temporally dislocated. Patcher rooms that have been created several months ago are still available now for users old and new to access and modify them. Kiwi stores collective patchers and never concludes them: on the network, they remain open for the community to keep modifying them. In this sense, collective events involving (or not) different places could also potentially be temporally disphased, and this is certainly a future development that needs to be taken into consideration.

Another important conceptual node is remoteness, which could be described as the simulation of a presence *in loco*, that manifests itself via patching operations on the computer screen. In considering systems of biometric detection, Joseph Pugliese borrows from Derrida the expression “metaphysics of presence” and identifies a series of potential failures involved in the technological exercise of recognising and passing on the indicators of human presence [16]. For the sake of this work, thus, it is useful to ask what are the limits of the simulation of presence in the context of live coding. More specifically, what fundamental losses relative to live patching are implied when users interact remotely? And, from a totally opposite perspective, does the possibility of

remote interaction make the choice of Kiwi over other patching programs meaningful and desirable only when remoteness is actually in place?

During the various remote sessions mentioned above, project members found themselves relying on textual comments on the patching canvas in order to make up for the absence of verbal communication with distant participants. The written text, in this case, operated to bridge a metaphysical gap in the interaction with other humans. This need for immediate verbal contact undoubtedly makes non-remote interactions meaningful and pertinent, alongside remote ones. In fact, again upon the April 11th server failure, the alternative session room opened by participants based in Brazil (available as ACRE-PARAÍBA TEST) soon ended up consisting in an interaction between two users that were working simultaneously on two different machines, occupying the same physical lab in Rio Branco, Acre. In this case, the interaction happened in the actual presence of the two participants, who had the possibility of establishing a higher level of empathy during the collaborative work, due to the possibility of immediate verbal and non-verbal communication. On top of being extremely meaningful when compared to remote interaction, this type of close and simultaneous interaction bridges an important gap in existing patching software: as mentioned above, in fact, collaborative live patching on Pure Data in the context of the project could only be consecutive, one-at-a-time interaction between users.

The geopolitical implications of remote live patching. As suggested above, relying on simultaneous interaction might be potentially problematic, as simultaneity demands high performance over a very specific and limited amount of time, and this might not be available to everyone, anywhere and at any given time. Geographical variation might also predict dramatic changes in connectivity, machine performance, etc. Difference in time zones might also be relevant, especially given the 7-hours gap between Rio Branco and Paris, João Pessoa being, at that time of the year, 2 hours ahead of Rio Branco and 5 hours behind Paris. These differences had a relevant influence on the practicality of the intercontinental sessions.

Although one of the principles that orientated the collaborative work was horizontality, the project ran the risk of aligning itself along very specific Cartesian coordinates of power and normativity [17]. However, given the commitment of the authors of this paper with forms of decolonial politics, this was aptly avoided. In fact, it is safe to affirm that the collaboration between the French and the Brazilian side of the project has remained horizontal.

Geopolitical or territorial conflict is a powerful metaphor that comes to mind when considering the concurrent activity of a relatively high number of agents on a patch canvas that is spatially limited. In this particular context, live patching does indeed become akin to fighting for a limited number of resources, or to fighting to secure some territorial control. Participants have the freedom to delete the objects or interrupt the connections created by others, and this tends to happen a lot depending on the number of simultaneously active users. The connection between the visual mapping of the GUI and the possibility of interaction of multiple concomitant users that characterises Kiwi permits to visualise the affinity between live patching as a musical activity and the politics of territorialised desire that inscribes human actions in space. In this sense, collaborative live patching is often the embodiment of some type of conflict, and it is possible to argue that this gives a particular, irrenunciable twist to the aural results.

Live patching, memory and the work on fragments. Obviously, the reference to conflict here is only symbolic and does not reflect the general climate of amity and comradeship that characterised the intercontinental sessions. These reflections, however, are useful in order to unearth the social and cultural implications of this type of activity. In this sense, another important "conflict" that characterised the first test session on April 11th was the divergence between a number of users that started from pre-patched, sizeable structures, and a number of users that attempted to build structures gradually, object by object. Soon after the test session, it was decided that pasting big structures was to be avoided, and that the live activity was to be implemented by working gradually.

Now, working gradually in real time with other users by interacting with, and at times by undoing, one's work involves activating memory in order to recall exact syntaxes, functionalities and even object names that might not always be ready by rote. In this context, the multiple dialectics between remembering and forgetting, and between retaining and effacing [18], produce a series of mnemonic fragments that play a crucial role in the dynamics of live patching. Participants end up relying on sketchy processes of recollection, rather than consolidating a steady corpus of program-specific knowledge. Here the aesthetics of the well-crafted, completed digital artefact, is supplanted by a poetics of the incomplete, of the experimental, of the open-ended.

6 Final remarks

The ongoing research reported in this paper permits the drafting of important conclusions and notable perspectives for the future collaboration between the groups involved in the intercontinental sessions.–The key node of the whole process is the meeting of local and global ecosystems. The patches LIVE-ACC-PATCH UFAC/UFPB and KiwiMaxPd2019 were developed locally by the Live/Acc/Patch group and the University Paris 8 class respectively. In this way, each ecosystem created specific patching strategies. During the first international session, these different strategies both met and clashed with each other, resulting in the emergence of a common collaborative practice, that culminated in the last session.

The complexity of the intercontinentally developed patches results from the combination of small elements that interact dynamically and turn into autonomous structures. With international live patching, the issue of temporality in simultaneous creation gains a new dimension. An entirely open, collaborative and non-hierarchical approach may be considered a downside by software developers. However, with Kiwi all the participants retain the same, unrestricted rights. In addition, the operations on each patch do not leave genealogical traces, that is, it is impossible to ascertain who created a specific object or comment on a patch. In this way, potential hierarchical barriers (i.e., lecturer vs. student or group X vs. group Y) are totally avoided. Kiwi and the know-how of live patching promoted by the Live/Acc/Patch group attest to an engagement with creative musical collaboration and computational development.

In this sense, this experience confirms the operativity of a common behavioural pattern in computer music history: the development of new tools is intrinsically connected to compositional and social issues that have specifically to do with musical

creation. In terms of the future developments of the intercontinental collaboration, two perspectives need to be highlighted. Firstly, in order to make up for Kiwi's technical limitations, it is possible to integrate the Faust language [19]. Secondly, From the point of view of musical performance and its morphological implications [20] live patching may perfectly be combined with experimental concert practices based on improvisation, open works and aural complexity.

References

1. Silva, L.A., Junior, J.E.F.N. Projeto Destino Pirilampo: um Estudo sobre a Composição de Meta-Soundscapes em Música Ubíqua. Revista Música Hodie 1(14), 105-121 (2014).
2. Keller, D., Barreiro, D. L., Queiroz, M., Pimenta, M. S. Anchoring in ubiquitous musical activities, in Proceedings of the International Computer Music Conference, University of Michigan Library, Ann Arbor, 319-326 (2010).
3. Keller, D., Lazzarini, V. ''Ecologically grounded creative practices in ubiquitous music,'' Organized Sound, 1(22) 61–72 (2017).
4. Musicoll homepage, http://musicoll.mshparisnord.org/, last accessed 2019/05/01.
5. CICM homepage, http://cicm.mshparisnord.org, last accessed 2019/05/01.
6. Ohm Force, https://www.ohmforce.com/HomePage.do, last accessed 2019/05/01.
7. Kiwi homepage, http://kiwi.mshparisnord.fr, last accessed 2019/05/01.
8. Puckette, M. Pure Data: another integrated computer music environment. In: Proceedings of the second intercollege computer music concerts, pp. 37-41(1996).
9. Magnusson T.: Algorithms as scores: Coding live music. Leonardo Music Journal (21), 19-23 (2011).
10. Collins N., McLean A., Rohrhuber J., Ward A., Live coding in laptop performance. Organised sound 8(3), 321-30 (2003).
11. Harazim, M. //This is a Comment: Music, Computers and Culture in Live Coding. Bmus dissertation, University of Aberdeen, 2017.
12. Pd~Graz homepage, https://pd-graz.mur.at/, last accessed 2019/05/02.
13. Ritsch, W. ICE - towards distributed networked computermusic ensemble. In: Georgaki, A., Kouroupetroglou, G. (eds.), Proceedings ICMC|SMC|2014, Athens, Greece (2014).
14. Chowning, J.M. The synthesis of complex audio spectra by means of frequency modulation. Journal of the audio engineering society 21(7), 526-534 (1973).
15. Paris, E., Millot, J., Guillot, P., Bonardi, A., Sèdes, A. Kiwi : vers un environnement de creation musicale temps réel collaboratif (premiers livrables du projet MUSICOLL). Journées d'Informatique Musicale 2017, Paris, France (2017).
16. Pugliese, J. The alleged liveness of "Live": Legal visuality, biometric liveness testing and the metaphysics of presence. In: Wagner, A., Sherwin. R. K. (eds.) Law, culture and visual studies, pp. 649-669. Springer, Dordrecht (2014).
17. Messina, M., de Araújo Souza, J. Rios, pontes, balsas e fronteiras: uma provocação desde a brasilidade liminar e precária do Vale do Rio Acre. Muiraquitã 6(1) 80-93 (2018).
18. Ricoeur P. Memory, history, forgetting. University of Chicago Press, Chicago (2004).
19. Faust homepage, https://faust.grame.fr, last accessed 2019/05/02.
20. Costa, V. F. Morfologia da Obra Aberta: esboço de uma teoria geral da forma musical. Editora Prismas, Curitiba, 2016.

Flexible interfaces: future developments for post-WIMP interfaces

Benjamin Bressolette and Michel Beaudouin-Lafon

LRI, Université Paris-Sud, CNRS,
Inria, Université Paris-Saclay
Orsay, France
benjamin.bressolette@lri.fr

Abstract. Most current interfaces on desktop computers, tablets or smartphones are based on visual information. In particular, graphical user interfaces on computers are based on files, folders, windows, and icons, manipulated on a virtual desktop. This article presents an ongoing project on multimodal interfaces, which may lead to promising improvements for computers' interfaces. These interfaces intend to ease the interaction with computers for blind or visually impaired people. We plan to interview blind or visually impaired users, to understand how they currently use computers, and how a new kind of flexible interface can be designed to meet their needs. We intend to apply our findings to the general public, to design adapted multimodal interfaces that do not always require visual attention.

Keywords: Multisensory interface, non-visual interface, flexible interface, blind users, visual attention, auditory, tactile.

1 Introduction

In Human-Computer Interactions (HCI), the large majority of information is transmitted to users through vision. This paper presents an ongoing project on interfaces that would be used without vision, by a combination of other modalities that convey rich and meaningful information. This process may lead to a new type of multimodal interfaces that we call *flexible interfaces*. These interfaces need to be flexible to users' desires, but also to users' abilities. In particular, we believe that visually impaired or blind users should more easily interact with computers. Digital tools enriched with new sensors and devices should be capable of going beyond people' disabilities.
In the next section, we briefly present related work on sound as information feedback, assistive technologies used by visually impaired and blind users, and interfaces that can inspire the design of a prototype. Section 3 develops the design process of a potential flexible interface adapted to blind and visually impaired users, before concluding in section 4.

2 Related Work

Computer interface mainly conveys visual feedback, thanks to the development of Graphical User Interfaces (GUI). Almost all recent computers have such an interface, and visual information is sometimes completed by informative sounds.

2.1 Sound as Information Feedback

Auditory icons introduced by Gaver [1] convey rich information and are still used nowadays, e.g. on Apple computers. The earcons are another example of icons, based on abstract sounds, and are widely used in computer interfaces [2]. Other sound icons has been designed, as spearcons that are based on accelerated speech [3], morphocons [4] or beacons [5].
Information can also be conveyed by continuous sounds through the sonification process [6]. Several studies showed how interactive sonification can help to improve movements [7, 8]. This continuous feedback is not used while interacting with computers, but it may be an interesting solution for this project.
As information are interactively transmitted through sounds, a design process is essential for aesthetic and emotional reasons. The Sonic Interaction Design in an interdisciplinary field that address the related interrogations, between "interaction design and sound and music computing" [9], and and this project can benefits from their work.

Even if sound icons are used on computers, information is mainly conveyed through vision, making these interfaces not adapted for visually impaired or blind users. Several systems has been designed to ease the transmission of visual information or to convert them into sounds or haptic stimuli.

2.2 Assistive Technologies for Blind or Visually Impaired Users

So-called WIMP – for Windows, Icons, Menus and Pointing – interfaces mainly give information through vision, leading to a serious problem of accessibility for visually impaired or blind people [10]. Some visually impaired people still rely on visual information, using for instance screen magnifiers [11]. However, palliative alternatives for blind users are based on visual information transmitted through another modality, namely sounds or tactile stimuli.

Screenreaders, which convert the visual information into speech, are broadly used [12–14]. These applications are useful for texts accessibility, but they are not really efficient with other contents, such as pictures or two dimensional representation and exploration. However, several improvements have been made to screenreaders, to address these issues [15, 16].
Braille interfaces are also widely used to convey textual information through touch [13, 17]. They are composed of matrix of pins that are controlled by the computer, as shown on Fig. 1. Several prototypes have been built with both devices to create rich multimodal interfaces [17–19].

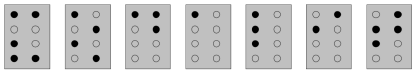


Fig. 1. An example of Braille interface, from Bellik and Burger [17].

These assistive technologies brings interesting alternatives for blind or visually impaired users. Nevertheless, several interfaces examples show principles that can help us reconsider actual GUI.

2.3 Beyond Actual Interfaces

An inspiring study was conducted by Newman and Wellner, with the ambition to reduce the incompatibilities between the physical and digital worlds [20]. They built an interface based on a desktop where real and digitial objects can be manipulated thanks to a videoprojector and a camera positioned above the desk. Recent prototypes have been built with the same principle [21, 22].
The previous example still involves a screen, but it is not the case of Gustafson et al.'s *Imaginary interface*, which is based on the relative position of the hands [23]. After specifying an L-shaped coordinate system with the non-dominant hand, users form shapes in the free space with the other hand. Even if visual attention is necessary for the coherence of the drawn shapes, this inspiring interface proposes fluid interactions without a screen.
In subsequent work, Gustafson et al. studied how the knowledge acquired while using a physical device as a phone can be transfered to build an *Imaginary Phone* [24]. As the users involuntarily learn the relative position of the icons of their phone, Gustafson proposed to use it remotely by pointing on the same relative position on their non-dominant hand. This study highlights how visual memory can contribute to design an interface manipulated with gestures and proprioceptive information, without requiring a screen. It adds an interesting flexibility to the classic smartphone, however vision is still needed to point to the desired icon.

These examples are still based on visual information, which are sometimes completed by sounds. We now give an example of a *flexible interface*, created for a specific situation where the user's sight is dedicated to a more important task.

2.4 MovEcho: an Interface Adapted to the Driving Context

MovEcho has been created to be used blindly, in a driving situation. Recent vehicles are equipped with touchscreens, but the major drawback of these interfaces is the visual distraction from the road they cause, which can lead to concentration problems and safety issues [25]. MovEcho has been designed to be manipulated with 3D gestures, and a rich sound feedback lets the user know if the gesture has been taken into account, which function has been manipulated,

and what is the current setting. The first version controls the ventilation system, and a second one manages the music system.
One of the major claim made is to use a virtual object as a natural link between the gestures and the sound feedback. This assertion is based on Gaver's work, describing the sound production in the environment as an interaction between two objects [26, 27]. In our opinion, creating a meaningful link between gestures and sounds had to connect the user's gestures performed in mid-air with an object, which we chose to add in virtual reality. Rath and Rochesso developed an idea about sonified virtual object that inspired this study [28]. The first theoretical results have been published [29], and the system has been compared to a touchscreen and tested in a driving simulator. The results show that even a first usage of MovEcho allows for a better visual task completion related to traffic, and is more appreciated by participants while driving.

Interfaces based on visual information should be more flexible with respect to users' abilities – e.g. is users' visual attention available – but also users' desires – e.g. do users want to use visual information. We agree with Dufresne on the need to develop "alternate modes of interaction for blinds or "sight occupied" users" [19], and that multimodal interfaces can be a promising solution.

3 Flexible Interfaces: Design Process and Future Work

In this project, we intent to rethink the construction of computer interfaces, starting by what is one of the foundations of these visual interface: the files, folders, and window management system.

3.1 Framework

The first step will consist of conveying rich information through the auditory and tactile channels, which can lead to the design of a first prototype. The different sequences of the project are detailed in Fig. 2, which is inspired by the framework described by Mackay and Fayard [30].

The literature provided examples of assistive technologies used by blind or visually impaired users, and inspiring ideas. However we believe that a first step in the design of a first prototype is to perform semi-structured interviews.

3.2 Semi-Structured Interviews

An important point is to understand how visually impaired or blind users interact with their computers, and in particular how they manage their files. How do they find a particular file on their computer? Do they use several windows? If so, how do they manage them? How do they manage the actual possibilities of their computer and the potential external devices to meet their needs? What are their strategies and workarounds to their impossibility to perceive visual information?

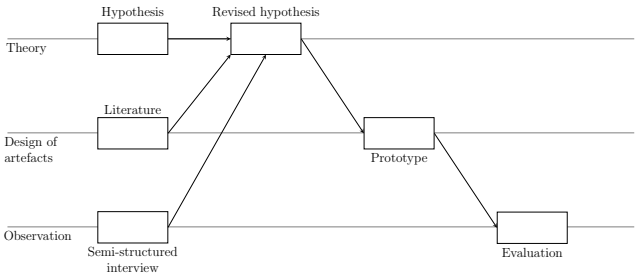


Fig. 2. Sequences of the project, with respect to the framework described by Mackay and Fayard [30].

To answer these questions we propose to conduct semi-structured interviews, during which the participants will be asked to perform live demonstrations while responding. The main questions of these interviews are prepared in advance, but the experimenter adapts his queries to the answers given, which explain why these interviews are qualified of "semi-structured". A common approach used is the critical incident technique originally proposed by Flanagan [31] and adapted to social sciences [32]: participants are asked to remember a recent incident, and give as much details as possible about it. The goal is to understand why this incident happened, why it was typical, and how it can be possible to avoid it in the future. This technique can be used on several topics, e.g. on files, folders and windows management in our case. Usually this procedure is very convenient before designing a prototype, as the answers can inspire design opportunities.

We plan to interview around 15 blind and visually impaired users, to understand the difficulties they can have while managing files, folders and windows. In particular, it would be interesting to interview participants not blind from birth, to figure out how they adapted, in particular while interacting with computers. These interviews are usually interpreted with a thematic analysis, which is formalized by Braun and Clarke [33]. The results would help to revise our initial hypothesis, see Fig. 2.

3.3 Prototype and future work

MovEcho is an example how post-WIMP interfaces can be designed, without only focusing on visual information. This flexibility would allow users to choose which modalities can be involved in the interacting process, and which have to be free for an other task, as traffic management or air traffic control [34]. For example in a face-to-face meeting, it would be better to dedicate vision to the interaction with the other person, and not to the computer.
As we mentioned, computers' interactions for blind users are only based on screenreaders and Braille interfaces. We advocate that substantial improvements

can be made with flexible interfaces, by directly designing multimodal interfaces rather than trying to convey visual information via another modality.

Multimodal objects can be the basis of the flexible interfaces we propose to build, e.g. with MovEcho. For instance, it may be interesting to extend to several modalities the notion of file, which actually can be manipulated with a mouse and give a feedback only by visual information. In our project, we may add the possibility to manipulate file objects with gestures, e.g. via tangible objects [35]. The system can give an information feedback to the user through sonification and haptic stimuli, this combination being interesting to avoid visual information [36]. The different ways to interact with the proposed multimodal objects may add this flexibility property to actual interfaces, that may be interesting for blind users or even sighted users that need vision for another task. In this spirit, MovEcho is a first step in this direction. Overall, it may be hard to manipulate all files and folders with this principle, but we think this path is worthwhile exploring. Finally, an evaluation on the field with blind and visually impaired users may follow the prototype design step, as detailed on Fig. 2.

4 Conclusion

This article addresses the limits about unimodal interfaces, which are mainly based on visual information and can even be dangerous to use while driving [25]. In our ongoing project, we propose to enrich these interfaces with stimuli from a variety of modalities to create flexible interfaces, which would allow users to interact while performing other actions.
These interfaces can also be flexible concerning the abilities of the users. In order to add rich information to existing interfaces, we propose to work with blind or visually impaired people. Interviewing these users with special needs can help us understand how they use current interfaces, their strategies and workarounds to interact without visual information. We also hope that these interviews will lead to design opportunities for the first prototype we propose to build. Afterwards, this interface can be tested in a field study, to identify the benefits and further improvements. We plan to adapt the findings of these studies to create multimodal interfaces for the general public.

References

1. Gaver, W. W.:Auditory icons: Using sound in computer interfaces. Human-computer interaction **2**(2), 167–177 (1986)
2. Blattner, M. M., Sumikawa, D. A., Greenberg, R. M.: Earcons and icons: Their structure and common design principles. Human–Computer Interaction, **1**(1), 11–44 (1989)
3. Walker, B. N., Nance, A., Lindsay, J.: Spearcons: Speech-based earcons improve navigation performance in auditory menus. In: Proceedings of the International Conference on Auditory Display, pp. 63–68. London, UK (2006)

4. Parseihian, G., Katz, B. F.: Morphocons: A new sonification concept based on morphological earcons. Journal of the Audio Engineering Society, **60**(6), 409–418 (2012)
5. Kramer, G.: Sonification system using auditory beacons as references for comparison and orientation in data. U.S. Patent No 5,371,854 (1994)
6. Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J. H., Miner, N., Neuhoff, J.: Sonification report: Status of the field and research agenda. In: Proceedings of the International Conference on Auditory Display, Santa Fe, USA (1999)
7. Schaffert, N., Mattes, K.: Acoustic feedback training in adaptive rowing. In: Proceedings of the International Conference on Auditory Display, pp. 83–88. Georgia Institute of Technology, Atlanta, USA (2012)
8. Danna, J., Velay, J.-L., Paz-Villagran, V., Capel, A., Petroz, C., Gondre, C., Thoret, E., Aramaki, M., Ystad, S., Kronland-Martinet, R.: Handwriting movement sonification for the rehabilitation of dysgraphia. In: Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR), pp. 200–208. Marseille, France (2013)
9. Serafin, S., Franinović, K., Hermann, T., Lemaitre, G., Rinott, M., Rocchesso, D.: Sonic interaction design, The Sonification Handbook, chapter 5, pp. 87–110. Logos Publishing House, Berlin, Germany (2011)
10. McGookin, D., Brewster, S., Jiang, W.: Investigating Touchscreen Accessibility for People with Visual Impairments. In: Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges, pp. 298–307. ACM Press, Lund, Sweden (2008)
11. Chiang, M. F., Cole, R. G., Gupta, S., Kaiser, G. E., Starren, J. B.: Computer and World Wide Web Accessibility by Visually Disabled Patients: Problems and Solutions. Survey of Ophthalmology, **50**(4), 394–405 (2005)
12. Watanabe, K., Shimojo, S.: When Sound Affects Vision: Effects of Auditory Grouping on Visual Motion Perception. Psychological Science **12**(2), 109–116 (2001)
13. Dufresne, A., Philippe Mabilleau, P.: Touching and Hearing GUI's: Design Issues for the PC-Access System. In: Proceedings of the Second Annual ACM Conference on Assistive Technologies, pp. 2–9. ACM Press, Vancouver, British Columbia, Canada (1996)
14. Kane, S. K., Bigham, J. P., Wobbrock, J. O.: Slide Rule: Making Mobile Touch Screens Accessible to Blind People Using Multi-Touch Interaction Techniques. In: Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 73–80. ACM Press, Halifax, Nova Scotia, Canada (2008)
15. Morris, M., R., Johnson, J., Bennett, C., L., Cutrell, E.: Rich Representations of Visual Content for Screen Reader Users. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 59. ACM Press, Montreal, QC, Canada (2018)
16. Meijer, P.B.L.:An experimental system for auditory image representations. IEEE Transactions on Biomedical Engineering **39**(2), 112–121 (1992)
17. Bellik, Y., Burger, D.: The Potential of Multimodal Interfaces for the Blind: An Exploratory Study. In: 18th Annual RESNA Conference (1995)
18. Brewster, S., Lumsden, J., Bell, M., Hall, M., Tasker, S.: Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices. In: Proceedings of the SIGCHI conference on Human factors in computing systems pp. 473–480. ACM, Ft. Lauderdale, Florida, (2003).
19. Dufresne, A., Martial, O., Ramstein, C.: Multimodal User Interface System for Blind and "Visually Occupied" Users: Ergonomic Evaluation of the Haptic and Auditive Dimensions. In: Human—Computer Interaction, pp. 163–68. Springer US, Boston, USA (1995)

20. Newman, W., Wellner, P.: A Desk Supporting Computer-Based Interaction with Paper Documents. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 587–592. ACM Press, Monterey, California, United States (1992)
21. Laviole, J., Hachet, M.:PapARt:interactive 3D graphics and multi-touch augmented paper for artistic creation. In: 2012 IEEE Symposium on 3D User Interfaces, pp. 3–6. Costa Mesa, CA, USA (2012)
22. Albouys-Perrois, J., Laviole, J., Briant, C., Brock, A.:Towards a multisensory augmented reality map for blind and low vision people: A participatory design approach. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 629. ACM Press, Montreal, QC, Canada (2018)
23. Gustafson, S., Bierwirth, D., Baudisch, P.: Imaginary Interfaces: Spatial Interaction with Empty Hands and without Visual Feedback. In: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, pp. 3–12. ACM Press, New York, USA (2010)
24. Gustafson, S., Holz, C., Baudisch, P.: Imaginary Phone: Learning Imaginary Interfaces by Transferring Spatial Memory from a Familiar Device. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 283-292. ACM Press, Santa Barbara, California, USA (2011)
25. Dingus, T. A., Klauer, S. G., Neale, V. L., Suzanne A. P., Lee, E. , Sudweeks, J. D. , Perez, M. A., Hankey, J., Ramsey, D. J., Gupta, S.: The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment. Technical report (2006)
26. Gaver, W. W.: How Do We Hear in the World? Explorations in Ecological Acoustics. Ecological Psychology **5**(4), 285–313 (1993)
27. Gaver, W. W.: What in the world do we hear?: An ecological approach to auditory event perception. Ecological psychology **5**(1), 1–29 (1993)
28. Rath, M., Rocchesso, D.: Continuous sonic feedback from a rolling ball. IEEE MultiMedia, **12**(2), 60–69 (2005)
29. Bressolette, B., Denjean, S., Roussarie, V., Ystad, S., Kronland-Martinet, R.: Harnessing Audio in Auto Control: The Challenge of Sonifying Virtual Objects for Gesture Control of Cars. IEEE Consumer Electronics Magazine **7**(2), 91–100 (2018)
30. Mackay, W. E., Fayard, A.-L.: HCI, natural science and design: a framework for triangulation across disciplines. In: Symposium on Designing Interactive Systems: Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques, vol. 18, pp. 223–234, (1997).
31. Flanagan, J. C. The critical incident technique. Psychological bulletin, **51**(4), 327–358 (1954)
32. Mackay, W. E.: Using video to support interaction design. DVD Tutorial, CHI **2**(5) (2002)
33. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative research in psychology **3**(2), 77–101 (2006)
34. Audry, E., Garcia, J.: Towards congruent cross-modal audio-visual alarms for supervision tasks. International Workshop on Haptic and Audio Interaction Design, Lille, France, (2019)
35. Baldwin, M. S., Hayes, G. R., Haimson, O. L., Mankoff, J., Hudson, S. E.: The tangible desktop: a multimodal approach to nonvisual computing. ACM Transactions on Accessible Computing (TACCESS), **10**(3), 9 (2017)
36. Hoggan, E. E.: Crossmodal audio and tactile interaction with mobile touchscreens. International Journal of Mobile Human Computer Interaction, **2**(4), 29–44 (2010)

Geysir: musical translation of geological noise

Christopher Luna-Mega[1] and Jon Gomez[2],

[1] University of Virginia, McIntire Department of Music
cjl9tx@virginia.edu
[2] University of Virginia, Data Science Institute
jag2j@virginia.edu

Abstract. The sounds of geological phenomena are generally noise. Wind, glaciers, oceans, streams, and other geological sounds present a vast content of frequencies that often obscures individual pitches or groups of pitches. However, noise varies from sound to sound with different pitch predominance and patterns. This variance contributes to the signature that makes several noise-sounds unique. In this study, the sound of one of the geysers in the Geysir system of the Haukadalur valley, 180 miles Northeast of Reykjavik, Iceland, is recorded and analyzed in multiple time segments, each with its own pitch predominance and, therefore, signature. The analysis is further adapted into a piece for seven spatialized pianists and electronics titled *Geysir*, which features the amplitude and predominant pitch class fluctuations throughout the geyser sample. This paper reports the process of the analysis and the compositional applications of the pitch class predominance analysis.

Keywords: Music Transcription, Mapping, Sonification, Ecoacoustics, Data Analysis, Algorithmic Composition, Music Information Retrieval

1 Introduction

Timbral analysis and musical translation of sound models derived from natural and human environments have established their place in instrumental music for the last fifty years. F.B. Mâche used the spectrogram to derive pitch information from analysis of a recorded sound in the early '60s [1]. "The train of thought he had elaborated became for many electroacoustic composers a conscious or unconscious aesthetic starting-point (e.g. musical landscapes and 'phonographies' in L. Ferrari's, M. Redolfi's, J.-C. Risset's and others' works)" [2]. He proposed "to bring together poetics and theory, and to show the advantages that there are in advancing an aesthetic project on the basis of a harmony with natural data" [3]. Since 1973, the French spectral composers worked with a similar "*ecological* approach to timbres, noises and intervals" [4]. A substantial part of their sound models generally derived from musical instruments with various degrees of harmonicity/inharmonicity in order to generate harmonic, motivic and structural foundations for instrumental music based on analyses of their harmonic spectra.

Geysir traces its lineage to three particular works from what Anderson calls the third phase of the spectral tradition [5], when spectral composers extended their territory to larger collections of sonic objects [6]: Murail's *Le partage des eaux* (1995) that

analyzes the sound of a wave crashing on the shore to generate material for orchestra [7]; *Bois Flotté* (1996), Murail's sequel to *Le partage,* that analyzes sounds of waves, swells and undertows to generate material for trombone, string trio, piano, and synthesized sounds [8]; and Ablinger's *Quadraturen IV* (1998), that analyzes a Berlin urban soundscape characterized by noise to generate material for a large ensemble [9].

In these works, as in *Geysir*, the analyzed sound model is geological noise. The higher degree of complexity in geological noise results in higher and more random quantities of data than that resulting from spectral analyses of harmonic sounds. In order to compose with geological noise sound models, it is necessary to simplify the mass of information. Fineberg explains how Murail achieved this in *Bois Flotté* through a reduction in the number of analysis-derived chords by re-sampling the sequence, decomposing it into narrow spectral slices, and quantizing the pitches [10].

Geysir adds to the contributions by preceding composers in its field by proposing a methodology that merges spectral analysis, statistical analysis, and performance indeterminacy. Like its predecessors, the analysis and compositional strategies for *Geysir* aim to embody the complexity of noise while at the same time simplifying it in order to reveal salient features of the sound model and make the music not exceedingly difficult to perform.

Re-synthesis derived partial tracking, explained below, was used for the initial stage. The re-synthesis process included octave segmentation and the deletion of partials below an amplitude threshold. Notation prototypes derived from the re-synthesis were generated in IRCAM's Open Music [11], yielding material of a high degree of complexity (see Fig. 1).

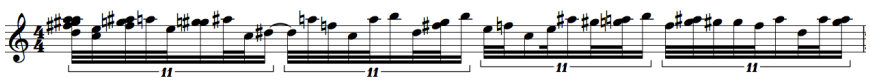


Fig. 1. staff 3 (C5-B5), 0:00-0:03, determinate partial tracking-derived notation

Considering the sound model's constant density and saturation from beginning to end, the performers would have had to sustain the rhythmic characteristics of this measure throughout the ~10-minute piece. From a perceptual perspective, the brevity of the rhythms paired to the saturation of the pitch sets was heard as a random mass of sound instead of auditory streams. The performance challenges of the material in addition to its aleatoric sound led to considering indeterminacy procedures for performance and rehearsal time economy. The indeterminacy consisted in having the performers generate the rhythmic material in a guided-improvisatory manner, with precise instructions regarding the pitch content, target rhythmic densities and dynamics. This idea led to a further distillation of the sound model: the calculation of pitch predominance at any desired point in time, organizing the pitches in three categories: high, medium and low predominance. The calculation was based on the total pitch count and durations per groups of measures, which provided the target rhythmic densities for each section. The pitch categories were assigned to specific note-heads (see Fig. 2).

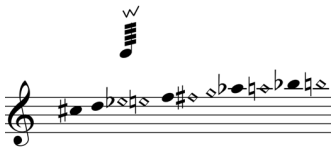


Fig. 2. staff 3 (C5-B5), 0:00-0:03, indeterminate pitch-predominance notation

Figure 1 features the first four seconds of the geyser in the frequency range between 523 and 1047 Hz (C5-B5) with determinate notation, while Figure 2 features an abstraction of the first 16 seconds of the same frequency range with indeterminate notation. The differences between the resulting music from the two notation versions were strikingly low. The sonic signature of the salient predominant pitches and the rhythmic density of the section were preserved in the indeterminate notation version, and the material was made more accessible for performers.

A significant byproduct of the pitch-predominance analysis and resulting indeterminate notation was the increased potential of the material to be embodied by the performers. Their agency with the indeterminate material established a closer connection to the sound model, and therefore to the geyser.

The following sub-sections will explain the processes for the geyser's pitch, rhythm and dynamics analyses and translation into music for seven pianists and electronics.[1]

2 Frequency Region Segmentation and Partial Tracking

Using the Sinusoidal Partial Editing Analysis and Resynthesis (SPEAR) software [12], the audio recording was resynthesized in order to manipulate, organize, and calculate the predominance of the frequency content (see Fig. 3).

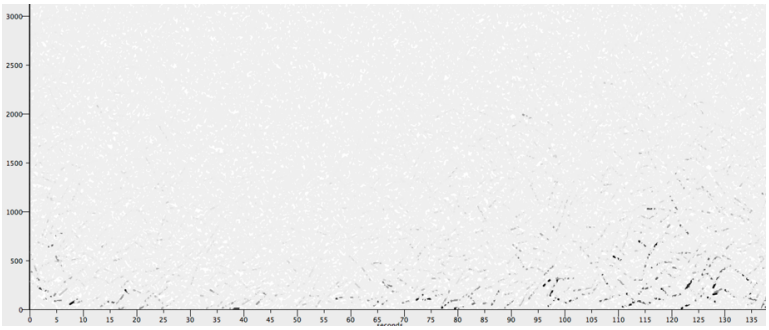


Fig. 3. Geyser analysis, entirety of re-synthesized partials until 2:15

[1] Full documentation of the analyses and audio examples referred to in the following sections available at http://www.christopherlunamega.com/works/analysis/geysir-analysis [13]

The audio was segmented in seven regions, from high to low, equivalent to a piano's seven complete octaves, from the lowest (C1) to the highest (C7). Each of these regions became an independent file (see Fig. 4). In each region, the partials with the amplitudes under -45 dB (the quietest) were eliminated, leaving only the loudest ones (see Fig. 5).

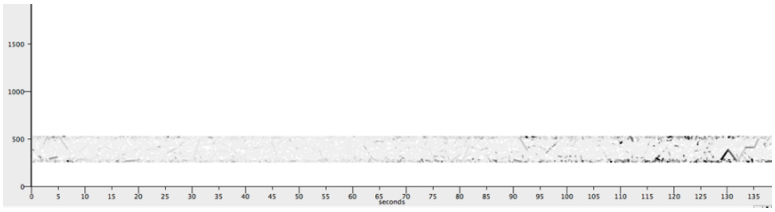


Fig. 4. Geyser analysis, segmented region (262-523 Hz)

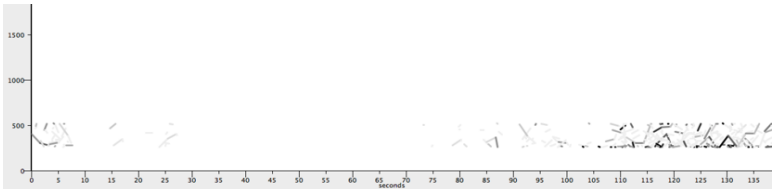


Fig. 5. Geyser analysis, segmented region after elimination of partials -25 dB

The sound data was converted from the SPEAR Sound Description Interchange Format (.sdif) into a text file (.txt) with IRCAM's Orchidée computer aided orchestration software [14]. Using Max MSP's Bach object library [15], the re-synthesized audio data encoded in the .txt file was then converted into music notation via partial tracking. A tool was generated for quantizing the complex rhythms and micro-tonal tunings of the geyser into simple rhythms adequate for pulse/time reference and the chromatic equal-tempered tuning system.[2] The resulting quantized audio data was translated into Music Exchangeable Markup Language format (.xml), making the contents compatible with data applications and music engraving software such as Sibelius or Finale.

[2] Equal-tempered tunings were chosen due to the fact that the music resulting from this analysis would be written for seven pianos. Quantization choices will vary depending on the affordances of the instruments that will perform. String instruments and some wind instruments can perform micro-tonal divisions up to 1/8 of a tone, as opposed to the piano, tuned to equal ½ of a tone.

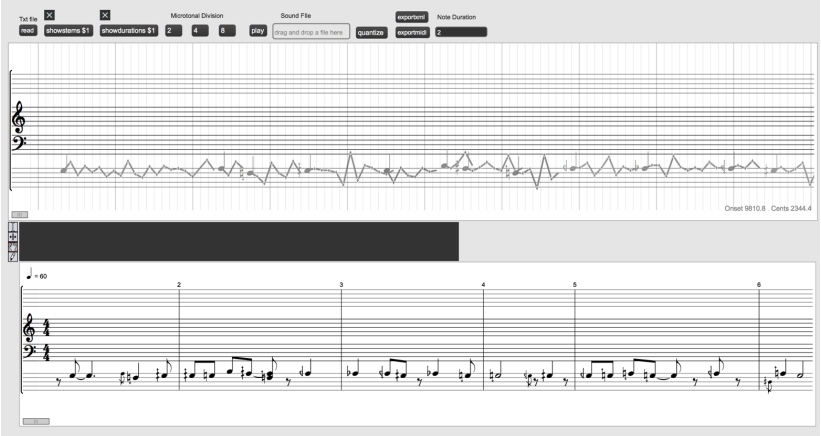


Fig. 6. Max MSP patch, including the bach.roll, bach.score, quantization, and .xml conversion tools[3]

3 Pitch Class Predominance Analysis

The last phase of the pitch analysis was the classification of the geyser's pitch classes derived from the previous processes. Using the BaseX database engine, all the pitches in the .xml file were organized by predominance and octave under the criteria of onset count and duration. A custom XQuery script [16] grouped the pitch classes and added the total time of their total onsets within a specific time segment (i.e., 16 onsets with a total of 20 seconds within a 25 second segment). Lastly, a list in descending order from most prominent to least prominent was generated for time intervals varying between 12 and 24 seconds. The lists derived from this analysis were then adapted entirely into the musical score for *Geysir*, for seven spatialized pianists and electronics.

3.1 Pitch Class Predominance Analysis Key

The information provided in this section is a description of each of the elements of the output from the XQuery script, which contains the distilled data used to inform the final scoring (see listing 1).

Listing 1. Pitch predominance data for staff 1 at 0:00-0:20

```
<group staff="1" measures="1,2,3,4">
 <pitch value="C#" duration="1920" count="15" rank3="high ●" rank4="high ●"/>
 <pitch value="C" duration="1408" count="11" rank3="high ●" rank4="medium ◇"/>
 <pitch value="Eb" duration="1280 count="10" rank3="medium ◇" rank4="medium ◇"/>
 <pitch value="D" duration="1280" count="10" rank3="medium ◇" rank4="medium ◇"/>
```

[3] The Max MSP programming was developed by Maxwell Tfirn [17].

```
<pitch value="F" duration="896" count="7" rank3="medium ◇" rank4="low ╳"/>
<pitch value="E" duration="256" count="2" rank3="low ╳" rank4="ruled out ∅"/>
```

Octave segmentation. The staves in the listing were numbered based on the initial octave segmentation performed in SPEAR.

- staff 1: C7 (piano 7) 2093-4186 Hz
- staff 2: C6 (piano 6) 1047-2093 Hz
- staff 3: C5 (piano 5) 523-1047 Hz
- staff 4: C4 (piano 4) 262-523 Hz
- staff 5: C3 (piano 3) 131-262 Hz
- staff 6: C2 (piano 2) 65-131 Hz
- staff 7: C1 (piano 1) 33-65 Hz

Pitch categories by predominance. The data shown in the listing displays pitches with calculated predominance in each octave, from highest to lowest, within the noise of the geyser. These pitches are labeled using four predominance categories: high (●), medium (◇), low (╳), and ruled out (⊘). These encodings were kept for the performers in the musical score.

Time. Time is presented in measures. Each measure is 4 seconds long, and the listing shows calculations over grouped measures. For example, in the first segment of the analysis (i.e., `<group staff="1" measures="1,2,3,4">`), each measure is 4 seconds long, so that the total time of measures 1, 2, 3, and 4 is 16 seconds. The temporal location of a measure is one less than the measure number multiplied by 4. For example, the time location of measure 30 is second 116, or time cue 01:56 (i.e., the first measure in `<group staff="1" measures="30,31,32,33,34">`).

Syntax. Each syntactic attribute in the document represents a specific feature used in the final scoring.

Table 1. Elements of the pitch predominance analysis

Feature	Description
Group staff	The octave analyzed (highest octave is group staff "1")
Measures	The total amount of measures in the segment analyzed
Pitch value	The pitch equivalence of the partial's frequency (e.g., 2218 HZ = C#)
Duration	Total duration of the partial's occurrences in the segment. The duration is displayed in milliseconds (1,920 milliseconds = 1.9 seconds)
Count	The number of iterations of the given pitch or frequency in the segment
Rank	The assessed predominance (low, medium, high). The "rank3" attribute is calculated from more pitches than "rank4", which filters out some pitches based on low counts/durations.

3.2 Pitch Class Predominance Analysis in Music Notation

The syntax of the analysis document (see Listing 2) was then converted into corresponding pitches on scored measures (see Fig. 8). The staff numbering translated into the register according to the octave segmentation shown above (staves 1-7 for C7-C1, respectively). The predominance ranks were mapped into note-heads with the same symbol (e.g., ● indicating a high predominance).

Listing 2. Pitch predominance data for staff 5 at 1:23-1:32

```
<group staff="5" measures="21,22,23,24">
 <pitch value="F#" duration="8320" count="64" rank3="high ●" rank4="high ●"/>
 <pitch value="Ab" duration="2816" count="21" rank3="low ╳" rank4="low ╳"/>
 <pitch value="G" duration="2560" count="20" rank3="low ╳" rank4="low ╳"/>
 <pitch value="E" duration="2176" count="17" rank3="low ╳" rank4="low ╳"/>
 <pitch value="F" duration="1408" count="10" rank3="low ╳" rank4="ruled out Ø"/>
 <pitch value="B" duration="1152" count="9" rank3="low ╳" rank4="ruled out Ø"/>
 <pitch value="C" duration="896" count="7" rank3="low ╳" rank4="ruled out Ø"/>
 <pitch value="Bb" duration="640" count="5" rank3="low ╳" rank4="ruled out Ø"/>
 <pitch value="A" duration="384" count="3" rank3="low ╳" rank4="ruled out Ø"/>
 <pitch value="D" duration="128" count="1" rank3="low ╳" rank4="ruled out Ø"/>
 <pitch value="C#" duration="128" count="1" rank3="low ╳" rank4="ruled out Ø"/>
```

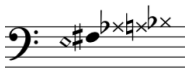
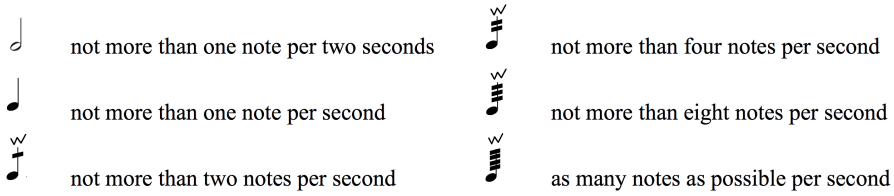


Fig. 8. Corresponding scoring for staff 5 at 1:20-1:36

4 Rhythmic Density Derived from Pitch Predominance

Six categories of rhythmic density are used in the scoring. The frame of reference is one second. The number of attacks per second defines the rhythmic category, as expressed below:



As mentioned in the introduction, the rhythmic density categories were derived from the pitch predominance analysis. The "duration" attributes in the pitch-classification listings in the previous sections (3.1, 3.2) informed the rhythmic density categories. Duration refers to the total time in milliseconds that the pitch is sounding in a segment of time.

The total duration was divided by the total number of seconds of the bars considered in the segment. For example, staff 1 presents C# as its most prominent pitch in measures "24, 25, 26, 27, 28, 29", with a total duration of 13,184 milliseconds (13.18 seconds) throughout the six measures. The duration per measure is 4 seconds. Therefore, the total duration of the 6-measure segment is 24 seconds. The 24 seconds of the segment divided by the 13.18 seconds in which C# is sounding results in an average rhythmic proportion of 1.8. Therefore, the rhythmic density for bars 24–29 is low, of not more than one note for every 2 seconds. At other points of the piece, the rhythmic density is quite high as the energy of the geyser and therefore its amplitude increases. In this sense, the amplitude contour of the sound model is generally connected to the rhythmic material of the piece.

As noise-derived harmony may result in what Grisey termed as "neutralization of pitch"[5] [4], the rhythmic density categories derived from pitch predominance, as well as spatialization, were implemented not only for the avoidance of monotony, but for perceptual clarity.

5 Amplitude Analysis

The seven re-synthesized frequency segments processed in SPEAR were imported individually to a Digital Audio Workstation. A waveform display was generated in which the y-axis represents amplitude in dB (decibels) and the x-axis represents time. A screenshot of each waveform display was segmented into seven dynamics regions: ***ppp***, ***pp***, ***p***, ***mp***, ***mf***, ***f***, ***ff***. A drawn contour was used to track the dynamic evolution of the geyser's frequency regions through time (see Fig. 9). Each of the frequency regions' contours was transcribed to each of the seven parts of the score.

The dynamic contours of the frequency regions with lowest amplitudes–staff 1 and staff 7, which present the highest and lowest frequencies–were occasionally altered for balance and intelligibility. For example, staff 1 presents a very brief peak at ***pp***, its highest amplitude in the entire 11 minutes of recording. For this reason, a sub-segmentation was made within the ***ppp*** range, where the highest peak is re-interpreted as an ***mp***. Figures of all staves are available in the *Amplitude contours by octave* pdf at http://www.christopherlunamega.com/works/analysis/geysir-analysis [13].

[5] Grisey was addressing the importance of new techniques that are necessary to avoid monotony.

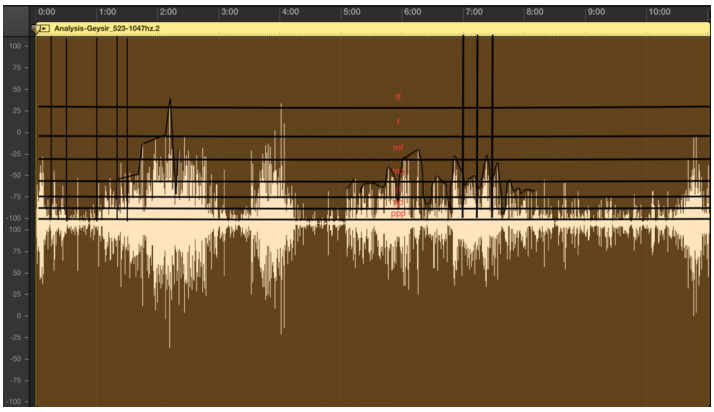


Fig. 9. Geyser analysis, staff=3 (C5–C6)

In the following example (see Fig. 10), the full score presents one of the overall peaks in amplitude in the entire 11-minute recording. A close look at the dynamics in each of the instruments will show the correspondences both in the macro-level and micro-level of dynamics: while there is a general increase in amplitude from 1:56 to 2:08, there are sudden dips and peaks in the dynamics within the overall increase in the section. The alterations in the dynamics of staff 1 and 7 are also evident in the example, in which the ***pp*** and ***p***, respectively, are increased to ***mp*** and ***mf*** in order to blend with the dynamics of the rest of the parts.

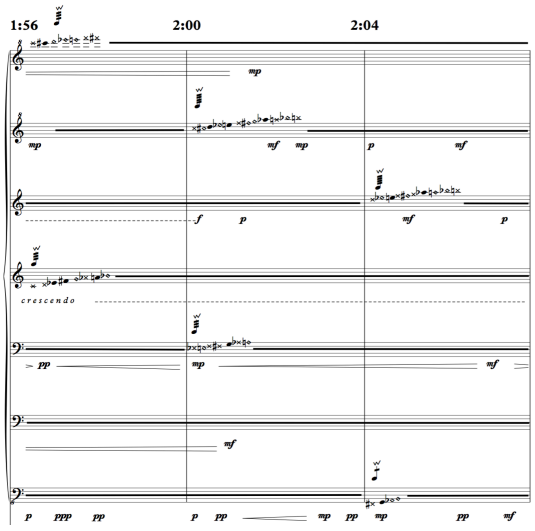


Fig. 10. *Geysir*, score excerpt

6 Notation

Each performer among the 7 pianists follows his/her part with a stopwatch. The pitch-predominance sets are introduced at varying intervals of time (between 12 and 30 seconds).

In the rhythm domain, the rhythmic value above the pitch-set determines the density that the performer will apply to perform the pitch set. The symbol placed above the rhythmic value means "irregular" (i.e., asymmetrical, uneven). The instruction to play irregularly is generalized in the entirety of the score, in accordance with the complexity of the sound model.

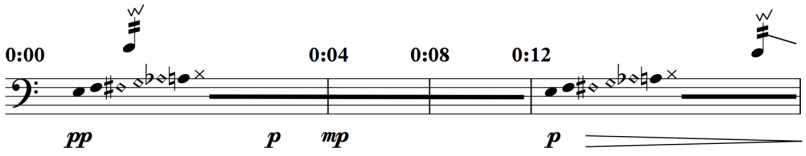


Fig. 11. *Geysir*, opening four measures, piano 5

The performer's choices are 1) the ordering of the sequences of the pitches included in the sets; 2) the durations of the irregular rhythmic values (in the example above, four notes per second, irregularly). The combination of these variables results in a variety of phrases that are generated by the performers, while preserving the essential features of the sound model (i.e., frequency, rhythmic density and amplitude contents).

7 Electronics / Spatialization

7.1 Spatialization

The electronics for the piece consist of seven-channel spatialized fixed media and telematic/live performer amplification. Due to the complexity of the sound model of the geyser and the music generated from it, spreading the streams of audio around the listeners was intended for perceptual clarity. The setup may vary depending on the venue, from a circular distribution of the speakers and all the performers (if the piece is performed in a flat level venue such as a museum), to a half circle distribution of the speakers, two performers on stage and five telematic performers. The latter version is the most viable for traditional concert halls and was the option for the premiere of the piece. The diagram for the setup is the following:

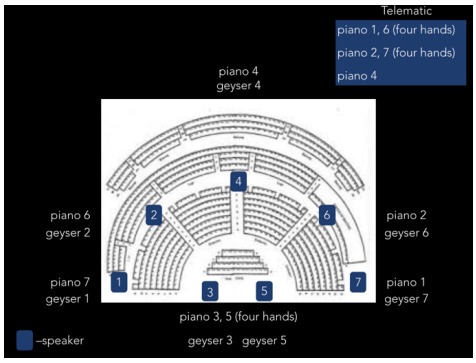


Fig. 12. *Geysir* spatialization diagram

Each speaker projects two sound sources: 1) telematic/live amplified piano; 2) a frequency stratus of the geyser sound model (explained below). There is one piano on stage, to be performed by two pianists playing piano 3 and piano 5, respectively. The other five pianists (pianos 1, 2, 4, 6, and 7) performed in piano cubicles situated outside the concert hall, in the University of Virginia's Department of Music. Their sound was sent to the concert hall using XLR cables and their image was broadcast live through a live video application projected on a screen on stage.

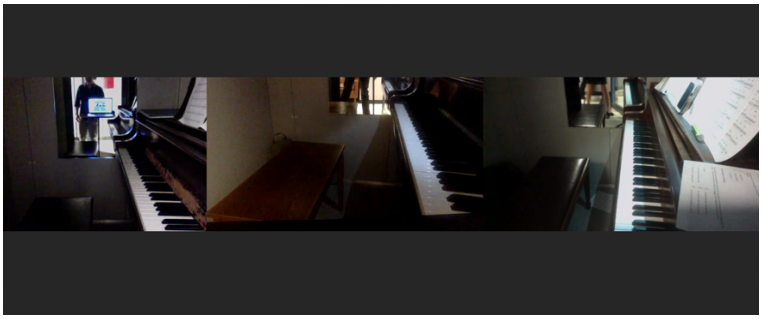


Fig. 13. *Geysir* telematic pianos

7.2 Fixed Media / Instrumental Pairing

Each of the seven tracks in the fixed media consists of a specific frequency stratum of the field recording of the geyser from which the piano parts are derived. Each track corresponds to the frequency range of each of the piano parts, based on their octave segmentation (see 3.1).

The spatialization diagram shows how each speaker projects two different audios containing different frequency strata. For example, Speaker 1 projects: 1) track 1 of the fixed media (the highest frequency region of the sound model) and 2) piano 7 (the lowest frequency region of the sound model-derived material). This approach was employed to balance the frequency distribution throughout the concert hall.

The score presents specific information regarding onsets/offsets of the tracks, as well as dynamics for both the fixed media and the live performers on the mixer.

7.3 Fixed Media–Live Instruments Amplitude Contour

The seven tracks in the fixed media do not sound simultaneously throughout the piece. Each track fades in and out of the mix in order to open the listening field to different frequency regions over time. When the general amplitude contour of the sound model is low, there is a small number of tracks active. Similarly, although not systematically, a larger number of tracks is active at the points of highest amplitude.

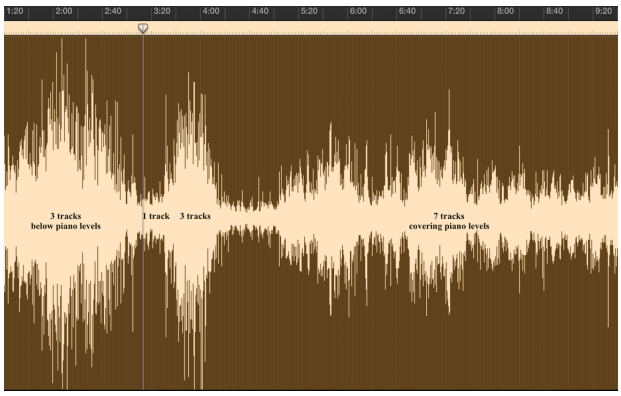


Fig. 14. Geyser analysis, general amplitude

The overall form of the piece has a general distribution of track density from lower to higher, all tracks being active for the last two minutes of the piece. The simultaneity of all tracks presents the sound model of the geyser at its full extent in an immersive spatialized environment. As this happens, the dynamics and amplification of the pianos subside, while the levels of the recording of the geyser in the fixed media are increased. By the last minute of the piece, the fixed media is the prevailing sound. The conceptual and poetic intention behind this formal plan was that the performers, who begin the piece with no electronics, gradually embody the sonic features of the geyser until they have become it. The electronics design is, in this sense, a representation of our philosophy of sound model-based composition: the embodiment of natural principles through the analysis of its sonic features.

8 Future Development

The modeling and scoring process could be improved by decreasing the required number of analytical manual steps. This could help musical authors develop scores based on similar phenomena with a broad spectrum of notes of highly-variable predominance. Other improvements could focus on the sound model itself, constructing a generalization of the sonic information to produce randomized scores with related sonic qualities. Finally, from a purely aesthetic point of view, a future compositional piece could explore less prominent frequencies.

References

1. O'Callaghan, J.: Spectral music and the appeal to nature. Twentieth-Century Music 15 (1) 57--73 (2018)
2. Grabócz, M.: The demiurge of sound and the poeta doctus. Contemporary Music Review 8 (1) 131--182 (1993)
3. Mâche, F.B.: Music, Myth and Nature. Hardwood Academic Publishers, Contemporary Music Studies 6, Chur (1992)
4. Grisey, G.: Did You Say Spectral? Contemporary Music Review 19 (3) 1--3 (2000)
5. Anderson, J.: A Provisional History of Spectral Music. Contemporary Music Review 19 (2) 7--22 (2000)
6. Malherbe, C.: Seeing Light as Color; Hearing Sounds as Timbre. Contemporary Music Review 19 (3) 15--27 (2000)
7. Murail, T.: Le Partage des eaux, http://www.tristanmurail.com/en/oeuvre-fiche.php?cotage=27533
8. Murail, T.: Bois flotté, http://www.tristanmurail.com/en/oeuvre-fiche.php?cotage=27532
9. Ablinger, P.: Quadraturen IV, http://ablinger.mur.at/docu11.html#qu4
10. Fineberg, J.: Musical Examples. Contemporary Music Review 19 (2) 115-- 134 (2000)
11. IRCAM Open Music Representation research group: Open Music, http://repmus.ircam.fr/openmusic/home
12. Klingbeil, M.: Software for Spectral Analysis, Editing and Resynthesis. In: ICMC (2005)
13. Luna-Mega, C.: Geysir, musical translation of geological noise, http://www.christopherlunamega.com/works/analysis/geysir-analysis
14. Carpentier, G., Tardieu, D.: Orchideé Automatic Orchestration Tool, http://repmus.ircam.fr/orchidee
15. Agostini, A., Ghisi, D.: A Max Library for Musical Notation and Computer-Aided Composition. Computer Music Journal 39 (2) 11—27 (2015)
16. World Wide Web Consortium: XQuery 3.1: An XML Query Language, https://www.w3.org/TR/2017/REC-xquery-31-20170321/
17. Tfirn, M.: Spectral Analysis Helper, https://maxwelltfirn.com/software-and-instruments/

Visual Representation of Musical Rhythm in Relation to Music Technology Interfaces - an Overview

Mattias Sköld

Royal College of Music in Stockholm
KTH Royal Institute of Technology
mattias.skold@kmh.se

Abstract. The present paper presents an overview of the ways we make sense of rhythm through visual means in music in terms of visual representation and notation, relating this to the user interfaces of music technology. Besides enabling the communication of rhythmical ideas, our systems of music representation reflect how we make sense of rhythm as a music parameter. Because of the complexity of visually representing rhythm, only software-based solutions provide flexible enough representations of rhythm in user interfaces. While the user interfaces of much rhythm oriented music technology deal with rhythm in looped phrases of 4/4 time, there are several examples of tools that challenge conventional ways of working with and visually representing rhythm.

Keywords: Music notation, rhythm, music technology interfaces

1 Introduction

My reasons for discussing the notation of rhythm in relation to music technology stem from my ongoing research exploring the possibilities of representing and notating pitch-based and sound-based music for composition [1, 2] with the aim of bridging the gap between acoustic and electroacoustic music. As a composer active in both fields, I am much aware of the differences in how rhythm is treated, produced and visualised in these fields. Also informing this paper is my background as a live-electronics musician where I have explored different ways of playing and interfacing with rhythm during live performances.

In his text on musical pulse perception, Parncutt defines musical rhythm as sequences of perceived events defined by their relative positions in time and their salience [3]. Traditional notation was developed to represent such sequences and their relative positions. But, as Edgard Varèse points out, new musical ideas call for new solutions for notation, particularly with regard to music technology [4]. Music theorist William Sethares divides notation into three categories: symbolic, literal and abstract, where literal systems such as waveform representations allow for the recreation of the music while symbolic approaches represent high level information. Abstract approaches are more of artefacts themselves or systems

limited to particular composers and their work [5]. Pierre Couprie, citing Jean-Yves Bosseur, talks of three axes of transmission: symbolic, graphic and verbal, the last two seeing an increase in use in music during the 20th Century [6]. Couprie adds algorithmic notation as a category, which refers to notation generated by coding, for example in software environments such as OpenMusic[1]. I would also include animated scores in this list since they relate to time differently from static scores. An important aspect of the visual representation of rhythm is the design of music technology interfaces for working with rhythm. Many creators of music today start from a position of doing rather than theorising, and not everyone working with electronic music knows music notation. In this text I will attempt to describe what defines the visual representations of rhythm that we work with in Western music today. I will exemplify the forms we use for visual rhythm representation for different musical purposes and how these forms and their features are represented in music technology.

1.1 Development and Forms of Visual Rhythm Representation

Marking stress patterns for the reading of poetry are among the earliest forms of rhythmic notation, being used by the ancient greeks [5]. Also non-Western cultures find origin of rhythmic notation in poetry e.g. Arabic music [7]. In the West, the 12th Century saw the development of square notation rhythmic modes and the mensural systems, providing advanced rhythmic possibilities [8]. Besides linear notation there are early examples of circular representations of rhythm, e.g. by Safī al-Din al-Urmawi, dividing a circle into 16 parts, starting from the twelve o'clock position and reading counter-clockwise [7].

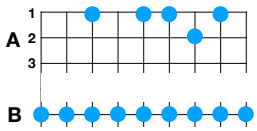


Fig. 1. The notation style of Xenakis' *Psappha* [9] where the letters are different instrument setups with different numbered instruments.

Five common forms of visual rhythm representation today are traditional music notation, circular, grid-based, numerical and waveform representation. Considering our preference for simple ratios in the perception of rhythm [10] it's understandable that traditional notation with its emphasis on duration relations remains the standard for musical rhythm representation. Circular representation has been used throughout history for repetitive and loop based rhythmic material. For endless loops, it makes sense to use modes of visual representation

[1] http://repmus.ircam.fr/openmusic/home

without start and end points. The distribution of drum tablature, as well as guitar and bass tablature, over the internet, makes good use of a type of grid-based representation. In this tablature, rhythm is indicated by means of monospaced font characters where each character represents one time unit. Numbers or letters mark the onsets of sounds and hyphens mark the absence of sound. Another example of grid-based rhythm representation is Iannis Xenakis' *Psappha* (1975) [9] for percussion solo, where the instrument onsets are indicated as dots on a grid (see Fig. 1). What characterises grid-based representations is the focus on note onsets and the distance between them. Music research and computer-assisted composition are two fields dealing with representation of rhythm as numbers. Examples of numerical representations of rhythm often have visual correlations as exemplified in Fig. 2, though the numerical representations can be seen as a visual representations in their own right. In terms of standards, MIDI data is the most common numerical representation of musical rhythm. It remains a widely used standard for representing musical data in all kinds of musical software. Waveform representations of rhythm are used in audio software and real-time composition environments where rhythm information needs to be deciphered from judging the distances between amplitude peaks along the time axis.

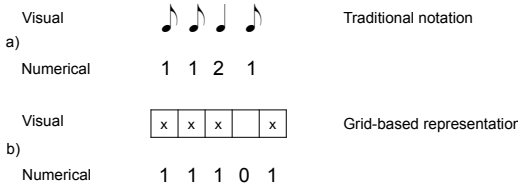


Fig. 2. Examples of numerical rhythm representation and corresponding visual representation

2 Features of Visual Rhythm Representation

2.1 Indicating Durations

The two main categories for visual rhythm representation are symbolic and graphic notation. Which one is used depends on the sound material. For continuous signals, graphic scores are suitable, while symbolic notation works best with separate sound objects. For visualising durations, one must first decide whether these should be expressed as successions of individual durations, as time indications along a timeline or as a combination of the two. Indicating *successive durations* is common for symbolic notation aimed for music performance. Durations indicated with regard to *elapsed time* is more common in electroacoustic music analysis and for the performance of pieces with little or no sense of pulse. There are also examples of hybrid approaches where music aimed at performance is notated as traditional relative pulse-based rhythm while also syncing

to a fixed media electroacoustic part or a movie using indications of absolute time such as SMPTE time code. While traditional notation can be used for relatively precise indications of note onsets and lengths by combining notes and rests, duration-related information may also be notated as articulation, e.g. *staccato* and *staccatissimo* articulations which result in shorter or much shorter sounds than notated.

2.2 Timelines

While the direction of the timeline for visual rhythm representation is almost exclusively left-to-right oriented there are exceptions such as circular visual representation or the top-to-bottom vertical flow of numerical rhythm data found in MIDI event lists. Also, while the timeline tends to be continuous, there are examples of musical works where fragments of music are displaced over the score to be performed in a random order, as in the case of Stockhausen's *Klavierstück XI* [11]. The timeline may indicate a) elapsed time or time code, b) measures and beats, or c) both at the same time. With regard to the last category, it's important to point out that these two timelines from a musical standpoint are conceptually different and there may be friction in applying the notion of absolute time to traditional notation of rhythm. When discussing the interpretation of traditional notation, expressive timing [12] needs to be considered. Performers of notated music typically enjoy an amount of freedom in the interpretation of the notated music which is reflected in the fact that the music sometimes has no tempo indication at all, or vague notions like *adagio* (slowly).

A decision with regard to the timeline, particularly when using graphical indicators of duration is whether to include a grid to clarify the progression of time. The inclusion of a grid as the background for rhythm visualization suggests positions one can align with or deviate from. Consider the difference in appearance of the examples in Fig. 3. What looks like a rhythm representations with no particular significance in example a), may seem like a rhythm failing to align with the beat structure of the piece in b), which has been "corrected" in example c). For tempo-based music such grids may be of help while for non-tempo-oriented music they can be very distracting.

While graphical indicators by definition adapt proportional graphic spacing, symbolic notation typically has the symbols placed closer together to save space. Though the symbols themselves provides sufficient information, proportional graphic spacing of notes much increases the readability of the rhythm. The Music Notation Project[2] recommends some degree of proportional spacing in their list of criteria for new notation systems.

With dynamic system such as animated scores, the timeline becomes subject to time itself making for interesting questions such as how much of the timeline is visible at any given point and how the passing of time is indicated as the score is animated.

[2] http://musicnotation.org/systems/criteria/

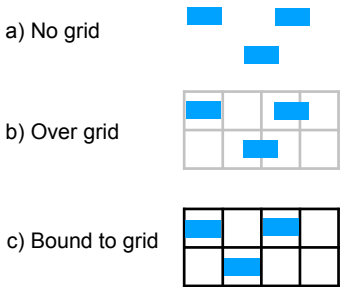


Fig. 3. Example of how the inclusion of a grid affects the visual appearance of rhythmic material.

2.3 Structural Hierarchies

Hierarchies in musical structures are a significant feature, not only in the functionality of traditional symbolic notation, but in many forms of music analysis, from Schenkerian analysis to Thoresen's auditive analysis of musical structures [13]. And though the term structure may sound like an objective feature, it's important to point out that it belongs to the perceived notions of music [14]. For an arbitrary sequence of note onsets different listeners may infer different tempos, and bar structures, if these are perceived at all. For algorithmic composition of instrumental music, hierarchical numerical structures is a convenient way of defining musical rhythm notation. Jacquemard, Ycart and Sakai [15] have explored using rhythm tree languages as a means for writing traditional music notation as writing language, useful for automated music transcription and computer-assisted composition (See Fig. 4).

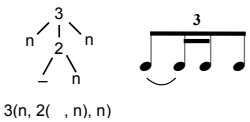


Fig. 4. Example of numerical rhythm representation and corresponding visual representation using a rhythm tree language [15]

3 New Tools for Visual Rhythm Representation

3.1 Tools for Analysis and Research

Depending on the rhythmic material, different non-linear models have been proposed for visualising the rhythmical structures. Fernando Benadon suggests using

circular plots to make sense of repetitive musical rhythms [16]. Toussiant explores how one can use geometry for understanding relations of rhythms plotted on circles [17]. Desain and Honing [18] have developed the concept chronotopological maps as a way of defining a three-interval rhythm as a single point in a triangle. This is done by indicating each of the three intervals' durations as values between 0-1 seconds along the three sides of a triangle. The points on each side of the triangle then point to one single position inside the triangle. This visual representation tool has been mainly used for rhythm perception research as in Jacoby and McDermott's studies on integer ratio priors on musical rhythm [10].

3.2 Dynamic Systems

Cat Hope, known for her work with animated scores, defines this category as "a predominately graphic music notation that engages the dynamic characteristics of screen media." And she argues that animated scores can be particularly useful for meeting the needs from aleatoric music and music for electronic instruments [19]. Animated scores bring a second layer of time since the animation happens in time, while the music proposed by the score happens in a separate time layer. The animation as well as the scored events may accelerate and slow down independently of one another. There is also the parameter of how much of the score a performer is allowed to see before playing, and whether there are interactive or generative elements in the creation of the animated score. Hope cites Gerhard E. Winkler placing animated scores somewhere between improvisation and fixed scores [19].

Algorithmic scores are scores based on underlying algorithms affecting their behaviour. Thanks to scoring tools like Bach [20] and MaxScore [21] it is now relatively easy to generate scores of traditional notation in Max in real-time paving the way for interaction and improivsation with scoring elements. These capabilities were explored by Pedro Louzeiro in the Comprovisador project for human and computer interaction with dynamic notation over a network of computers [22].

Thor Magnusson speaks in [23] of live coding algorithms as scores. The instructions to the computer acting as a form of score in its own right for the interpreting computer to perform. Magnusson relates this category to previous work in contemporary music with open forms on the one hand, and music machine scores, such as player piano rolls, on the other. It is also worth pointing out that the text-based computer music language Csound uses the terms score and orchestra for the two files defining performance and synthesis. It is easy to see the parallel between instructions for a performer and instructions for a computer, though it is debatable whether a computer algorithm defining musical rhythm behaviour can be thought of as a visual representation of said rhythm. Also, different programming languages and environments have different levels of visual clarity. Max, being a visual programming language, can make structural components of a rhythmical structure apparent in how the patch cords are drawn and how subpatches are named and annotated.

3.3 Three-Dimensional Scores

Expanding on the dimensions used for scoring musical parameters has opened up new possibilities for thinking of visualised music. The idea is not new - Earle Brown's *December 1952* [24] is a graphical score that can be read from any direction. The score can be thought of as a two-dimensional map rather than a linear string of parameter values. David Kim-Boyle has worked with Jitter[3] to produce graphical scores, where performers trace a path through a system of nodes positioned in a 3-D space [25].

Andrew McGraw has explored the potential in using three-dimensional visual representation for music analysis, showing how a repetitive musical structure can be viewed as a three-dimensional helix (see Fig. 5 for an example of helical representation), drawing on the potential of displaying repetitive music in circles, while accounting for changes over time [26].

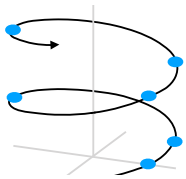


Fig. 5. Example of helical representation of rhythm.

4 Visual Rhythm Representation in Music Technology Interfaces

The two common technologies for work with rhythm in music technology are sample buffers with recorded rhythmic material and event sequencers triggering each sound. Sample buffers, when visualised, are usually displayed using continuous waveform representation, while event sequencers tend to work with various forms of symbolic representation. Durations are usually divided into two categories: the delta time between onsets of sounds and the time lengths of each sound. It is worth noting that user interfaces of rhythm sequencers generally only treat the delta time between onsets as a musical parameter while the actual duration is often a parameter of sound production, defined by amplitude envelopes. This echoes the division in traditional notation between rhythm notation and articulation mentioned above. While providing more information, waveform representation is also most suitable for visualising distances between sound onsets of rhythms.

The timeline for a hardware sequencer is usually a fixed number of steps to be triggered at a certain tempo. Computer sequencers in DAWs like Logic

[3] The graphical functionality of Max

Pro can work with timelines both with absolute time and pulse-based time. Therefor there are functions for modifying material in one domain to fit the other. Ableton Live[4] stands out, however, since its main window is not the timeline, but a loop mixer environment with circular displays for loop progress. Linear and cyclic perspectives can be changed in modern DAWs, e.g. Logic Pro allows for the drum loop programming inside the UltraBeat plugin to be moved into the traditional arranger timeline, while any region can be turned into a loop by pulling a handle. Despite tempo having a crucial impact on a sequencer timelines this is not necessarily visualised. This is different with DJ-oriented technology such as Traktor Pro 3[5] since tempo matching and tempo-dependent effects are important.

Hierarchies are often visualised in event sequencers e.g. as dividing lines for every four steps to separate each 16th-note group. Hierarchies are also present in most software aimed for algorithmic notation, e.g. in the Bach library for Max, rhythms are displayed as lisp-style nested lists. Also, different kinds of rhythmical hierarchies are conveyed by the layout of a modular patch or the disposition of routines in live-coding.

4.1 New forms of Visual Rhythm Representation in Music Technology Interfaces

A standard modular sequencer is already a very flexible tool, but multi-dimensional sequencers take sequencing one step further by introducing the possibility of advancing a sequence in more than one direction, in a way acting as the sequencing counterpart to Brown's *December 1952*. The MakeNoise René[6] and TipTop Audio's Z8000[7] are examples of such sequencers with input for progression along the X as well as the Y axes.

Besides Ableton Live, there is music technology that have adapted the idea of the loop as a circle, perhaps most notably the sequencers Model 252e Polyphonic Rhythm Generator and Model 250e Dual Arbitrary Function Generator by Buchla[8] (there are also Eurorack Sequencer Modules with this approach). For real-time exploring of repetitive rhythmic structures it makes sense to have a visual representation approach without beginning or end points.

Godfried Toussaint's findings that a great variety of rhythms can be generated using the euclidean algorithm of computing the greatest common divisor [27], has been put to use in several sequencers, such as the vpme.de Euclidean Circles[9]. This marks an alternative way of generating rhythmic material. Euclidean sequencers distribute an arbitrary number of beats evenly over a grid-based phrase fixed length, making it possible to generate rhythmical patterns by

[4] https://www.ableton.com/en/live/
[5] https://www.native-instruments.com/en/products/traktor/dj-software/traktor-pro-3/
[6] http://makenoisemusic.com/modules/rene
[7] http://tiptopaudio.com/z8000/
[8] https://buchla.com
[9] http://vpme.de/euclidean-circles/

turning a knob. The position of the knob thus representing the event density of the rhythm.

4.2 Dynamic Systems

For working with dynamic possibilities of visual representation and time-based control of music there are a few systems designed to do just that. IanniX [28], an ambitious open source continuation of Xenakis pioneering graphical computer music project UPIC (Unité Polygogique Informatique du CEMAMu), enables flexible visual representation and control of music parameters by freely placing and configuring objects and trajectories to govern any kind of musical parameter. Timed Sequences [29] is similarly a framework for visualising and controlling musical events in time, working with musical structures as "timed items".

For animated scores, the Decibel Scoreplayer [30], associated with the Decibel New Music Ensemble[10], is an established system, though composers in this field also program their own solutions, like Ryan Ross Smith who developed the animated score for his Study no 50 [31] in openFrameworks. In a way, multi-page grooveboxes such as the Korg Electribe[11] can also form a sort of animated score, when the different pattern pages are displayed as they are performed, with flashing LEDs indicating sequencer progress.

Besides the dynamic systems mentioned above, much of the experimental work in visual rhythm representation is carried out in music programming languages such as Max[12], SuperCollider[13] and Chuck[14]. In fact, my own reason for first using Max was a rhythmical idea demanding multiple varying tempi. Much unconventional work with rhythm still demands tailor made solutions both with regard to sound production and visual representation. Music programming solutions have also been used for network distribution of notation e.g. using MaxScore [21].

4.3 Other Experimental Technologies for Musical Rhythm

It is worth mentioning that besides the highly flexible worlds of music programming and modular synthesis, there is software that tries new approaches to rhythm work and display in the form of tablet apps or DAW-plugins. Christian Gwiozda's experimental music sequencer Gridcomposer[15], Scott Garner's three-dimensional node-based sequencer Ostinato[16] and Kymatica's stochastic sample-based sequencer Sector[17] with time-warp functionality are three apps

[10] http://www.decibelnewmusic.com/
[11] https://www.korg.com/us/products/dj/electribe/
[12] https://cycling74.com/products/max/
[13] https://supercollider.github.io
[14] https://chuck.cs.princeton.edu
[15] http://www.gridcomposer.net
[16] http://www.scottmadethis.net/interactive/ostinato/
[17] http://kymatica.com/apps/sector

with novel approaches to rhythm sequencer functionality. Audio Damage's non-linear "neuron sequencer" Axon[18] and the Native Instruments' "Life MIDI-sequencer" Newscool[19] are two interesting software plugins where the Axon relies on a web of six hexagons as its user interface while the Newscool is based on the Life model developed by John Conway.

5 Discussion and Conclusions

Rhythm in electroacoustic music is fundamentally different from traditional acoustic music in that materials and tools in both Musique Concrète and electronic music themselves contain rhythm. The rhythm loops of the trains in *Étude au chemins de fer* by Pierre Schaeffer originated in the recorded sound material and were not preceded by a score. And the rhythms of Morton Subotnick's *Silver Apples of the Moon* were not written with symbolic notation, but were produced through work with modular synthesis. Schaeffer highlights this feature of *new music* (musique concrète) going from material, through a draft stage to a composition rather then the other way around as in *ordinary music* [32]. This means that music technology has a strong influence on the music made from it. The fact that instruments of music technology themselves produce rhythm poses interesting questions particularly when combining them with traditional acoustic instruments where recorded or stored music material was traditionally confined to music box cylinders and player piano rolls. Electronic instruments with built-in sequencers also tend to have the possibility of serving as masters or slaves in syncing to other instruments which adds a hierarchial layer to the rhythm work, particularly in collaborative performances.

At the heart of rhythm control in music technology is the sequencer. In the most general terms a sequencer is a bank of individual data set(s) that can be accessed in real time. The flexibility and layout of a sequencer depends on its intended use: sequencers in drum machines are typically grid-based event sequencers run by a master clock, while sequencers for modular systems are open for more experimental use. Therefore, in the case that a sequencer is used for rhythm, what is displayed by its indicators and knobs could at best be seen as rhythmical material rather than the rhythm itself.

When studying the functionality and visual interfaces of music technology aimed for musical rhythm it becomes clear that there are two major categories. On the one hand there are drum machines, grooveboxes and DAWs that may be flexible but were designed within the context of traditional music notation. Loops and sequences are easily placed in line with an underlying pulse or adapted to sync with other loops in the same machine or a different machine. When used as intended these machines usually provide relatively clear readings of the rhythm structure in play. On the other hand there are experimental systems like IanniX, that take little for granted both regarding visual representation and control.

[18] https://www.audiodamage.com

[19] Part of Reactor - https://www.native-instruments.com/en/products/komplete/synths/reaktor-6/

One might add a third category of technologies that take existing ideas and concepts and make them more flexible. MaxScore and Bach are libraries for traditional notation, but provide new functionality and behaviour to this category. Bringing music programming functionality to a DAW as in Max for Live[20], may also belong to this category. Needless to say, maximum flexibility in software, concerning both the functional and the visual, is found in the music programming languages themselves while flexibility in hardware is found in the world of modular synthesis. In both these cases it is up to the user to make the programming or patching (including positioning of modules) visually representative of the sounding structure.

References

1. Mattias Sköld. The Harmony of Noise: Constructing a Unified System for Representation of Pitch, Noise and Spatialization. In *CMMR2017 13th International Symposium on Computer Music Multidisciplinary Research*, pages 550–555. Les éditions de PRISM, 2017.
2. Mattias Sköld. Combining Sound- and Pitch-Based Notation for Teaching and Composition. In *TENOR'18 – Fourth International Conference on Technologies for Music Notation and Representation*, pages 1–6, 2018.
3. Richard Parncutt. The perception of pulse in musical rhythm. *Action and Perception in Rhythm and Music*, 55:127–138, 1987.
4. Edgard Varèse and Chou Wen-Chung. The liberation of sound. *Perspectives of new music*, pages 11–19, 1966.
5. William Arthur Sethares. *Rhythm and transforms*. Springer Science & Business Media, 2007.
6. Pierre Couprie. Algorithmique et technologies numériques dans la notation musicale. In *Musiques Orales, leur Notation et Encodage Numérique (MEI)*, pages 99–115. Les éditions de l'immatériel, 2015.
7. Habib Yammine. L'évolution de la notation rythmique dans la musique arabe du ixe à la fin du xxe siècle. *Cahiers d'ethnomusicologie. Anciennement Cahiers de musiques traditionnelles*, (12):95–121, 1999.
8. Ian D. Bent, David W. Hughes, Robert C. Provine, Richard Rastall, Anne Kilmer, David Hiley, Janka Szendrei, Thomas B. Payne, Margaret Bent, and Geoffrey Chew. Notation, 01 2001.
9. Iannis Xenakis. *Psappha*. Salabert, EAS 17346, Paris, 1975.
10. Nori Jacoby and Josh H McDermott. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3):359–370, 2017.
11. Karlheinz Stockhausen. *Klavierstück XI*. Universal Edition, Vienna, 1956.
12. Neil Todd. A model of expressive timing in tonal music. *Music Perception: An Interdisciplinary Journal*, 3(1):33–57, 1985.
13. Lasse Thoresen. Auditive analysis of musical structures. a summary of analytical terms, graphical signs and definitions. In *ICEM Conference on Electro-acoustic Music*, pages 65–90, 1985.
14. Henkjan Honing. Structure and interpretation of rhythm in music. *Psychology of music*, pages 369–404, 2013.

[20] https://www.ableton.com/en/live/max-for-live/

15. Florent Jacquemard, Adrien Ycart, and Masahiko Sakai. Generating equivalent rhythmic notations based on rhythm tree languages. In *Third International Conference on Technologies for Music Notation and Representation (TENOR)*, 2017.
16. Fernando Benadon. A circular plot for rhythm visualization and analysis. *Music Theory Online*, 13(3), 2007.
17. Godfried Toussaint. The geometry of musical rhythm. In *Japanese Conference on Discrete and Computational Geometry*, pages 198–212. Springer, 2004.
18. Peter Desain and Henkjan Honing. The formation of rhythmic categories and metric priming. *Perception*, 32(3):341–365, 2003.
19. Cat Hope. Electronic scores for music: The possibilities of animated notation. *Computer Music Journal*, 41(3):21–35, 2017.
20. Andrea Agostini. *Bach: An environment for computer-aided composition in max*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2012.
21. Georg Hajdu and Nick Didkovsky. Maxscore: Current state of the art. In *ICMC*, pages 9–15, 2012.
22. Pedro Louzeiro. Real-time compositional procedures for mediated soloist-ensemble interaction: the comprovisador. In *International Conference on Mathematics and Computation in Music*, pages 117–131. Springer, 2017.
23. Thor Magnusson. Algorithms as scores: Coding live music. *Leonardo Music Journal*, pages 19–23, 2011.
24. Earle Brown. December 1952. *American Music*, 26(1), 1952.
25. David Kim-Boyle. The 3-d score. In *TENOR 2017: International Conference on Technologies for Music Notation and Representation:[24-26 May 2017, University of A Coruña, Spain]*, pages 33–38. Facultade de Filoloxía, 2017.
26. Andrew McGraw. Preliminary remarks on the helical representation of musical time. *Analytical Approaches to World Music*, 3(1), 2013.
27. Godfried Toussaint. The euclidean algorithm generates traditional musical rhythms. In Reza Sarhangi and Robert V. Moody, editors, *Renaissance Banff: Mathematics, Music, Art, Culture*, pages 47–56, Southwestern College, Winfield, Kansas, 2005. Bridges Conference. Available online at `http://archive.bridgesmathart.org/2005/bridges2005-47.html`.
28. Guillaume Jacquemin, Thierry Coduys, and Matthieu Ranc. Iannix 0.8. *Actes des Journées d'Informatique Musicale (JIM 2012)*, pages 107–115, 2012.
29. Jérémie Garcia, Dimitri Bouche, and Jean Bresson. Timed sequences: A framework for computer-aided composition with temporal structures. In *Third International Conference on Technologies for Music Notation and Representation (TENOR 2017)*, 2017.
30. Cat Hope and R Lindsay. The decibel scoreplayer-a digital tool for reading graphic notation. 2015.
31. Ryan Ross Smith. [study no. 50][notational becoming][speculations]. In Richard Hoadley, Chris Nash, and Dominique Fober, editors, *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'16*, pages 98–104, Cambridge, UK, 2016. Anglia Ruskin University.
32. Pierre Schaeffer. *In search of a concrete music*. University of California Press, Trans. C. North, J. Dack. Berkley, Los Angeles, 2012.

A Tree Based Language for Music Score Description.

D. Fober[1], Y. Orlarey[1], S. Letz[1], and R. Michon[1]

Grame CNCM Lyon - France
{fober,orlarey,letz,michon}@grame.fr

Abstract. The presented work is part of the INScore project, an environment for the design of augmented interactive music scores, oriented towards unconventional uses of music notation and representation, including real-time symbolic notation capabilities. This environment is fully controllable using Open Sound Control [OSC] messages. INScore scripting language is an extended textual version of OSC messages that allows you to design scores in a modular and incremental way. This article presents a major revision of this language, based on the description and manipulation of trees.

Keywords: Music notation · Programming language · INScore.

1 Introduction

There is a large number of musical score description languages (Lilypond [11], Guido [9], MuseData [8], MEI [12], MusicXML [7] etc.) that are all turned towards common western music notation. The extension of some of these languages has been considered, in order to add *programmability* e.g. operations to compose musical scores in Guido [5], or the Scheme language in Lilypond. There are also programming languages dedicated to common music notation, like CMN [13] or ENP [10] that are actually Lisp dialects.

The approach proposed by INScore [4] is different: symbolic music notation is supported (via the Guido language and rendering engine [2,9]), but it constitutes one of the means of music representation among others, without being privileged. Purely graphic scores can be designed. All the elements of a score (including purely graphical elements) have a temporal dimension (date, duration and tempo) and can be manipulated both in the graphic and time space. The notion of time is both event-driven and continuous [6], which makes it possible to design interactive and dynamic scores. Figure 1 presents an example of a score realised using INScore. It includes symbolic notation, pictures, a video, and cursors (the video is one of them) which positions are synchronised by the performer gestures.

INScore has been initially designed to be driven by OSC messages [14]. OSC is basically a communication protocol. A textual version of the OSC messages constitutes the INScore storage format, which has been extended to a scripting language, [3] allowing greater flexibility in music scores design. These extensions (variables, extended addresses, Javascript section, etc.) have nevertheless suffered from a rigidity inherent to an ad hoc and incremental design. For example, the parser makes a clear distinction between OSC addresses and associated data, which prevents the use of variables in OSC addresses. Thus, a major revision of this language became necessary. It is based on the

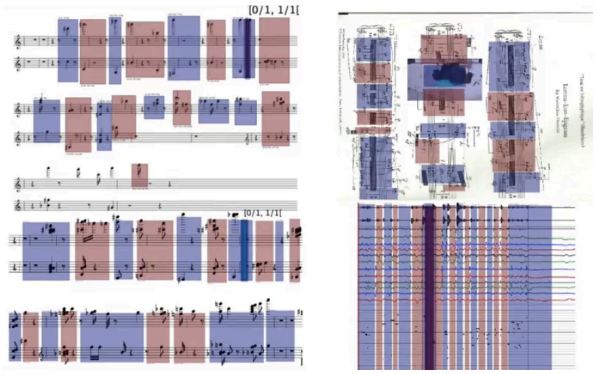


Fig. 1. A score realised using INScore, used as part of a sensor-based environment for the processing of complex music called *GesTCom* (*Gesture Cutting through Textual Complexity*) [1].

manipulation of a regular tree structure that is also homogeneous to the INScore model. Figure 2 gives an example of such model hierarchy, that can be described in the current scripting language (i.e. OSC) by listing all branches from the root.

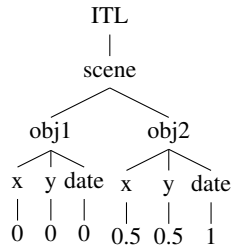


Fig. 2. A score sample including 2 objects $obj1$ and $obj2$, having x, y, and $date$ attributes. Time properties (date, duration) are notably used to represent the time relationship between objects.

After some definitions, we will present the basic operations on trees and the corresponding grammar.Then we introduce mathematical operations on trees, the concepts of *variables* and *nodes in intention* and we'll present how this language is turned into OSC messages. The final section gives an example of the new language before concluding.

2 Definitions

A tree t consists of a value v (of some data type) and the (possibly empty) list of its subtrees.

$$t : v \times [t_1, ..., t_k]$$

A tree with an empty list of subtrees $t : v \times [\,]$ is called a leaf.
A value is among literal (i.e., text, number) or special values of the following types:

- forest ($\varnothing$): denotes a tree including only subtrees,
- mathematical operators: indicates a mathematical operation between subtrees,
- variable: denotes a tree whose value refers to another tree,
- expand: indicates a tree to be expanded,
- slash (/): used for conversion to OSC

Use and evaluation of these values is detailed in the next sections.

3 Operations on Trees

We define two abstract operations on trees: sequencing and paralleling. These operations have no musical semantics, neither from a temporal nor from a graphic point of view. They are defined as methods for algorithmic construction of OSC messages and operate on the topological organisation of the trees.

3.1 Putting Trees in Sequence

Putting two trees t and t' in sequence adds t' as child of all the leaves of t. We will note $|$ the sequencing operation. Let 2 trees $t : v \times [t_1, ..., t_k]$ and t'. Then:

$$v \times [t_1, ..., t_k]|\ t' \rightarrow v \times [t_1|\ t', \ldots, t_k|\ t']$$

with:

$$\begin{cases} v \times [\,]|\ t' \rightarrow v \times [t'] \\ v \times [\,]|\ \varnothing \times [t_1, ..., t_k] \rightarrow v \times [t_1, ..., t_k] \end{cases}$$

The right arrow ($\rightarrow$) indicates the result of an expression evaluation.

3.2 Putting Trees in Parallel

Putting two trees t and t' in parallel consists in putting them in a forest. We will note $\|$ the parallelisation operation. Let 2 trees t and t':

$$t \parallel\ t' \rightarrow \varnothing \times [t, t']$$

The result is a tree which value $\varnothing$ denotes a forest.
Parallelisation applied to a *forest* preserves the subtrees order:

$$\begin{cases} \varnothing \times [t_1, ..., t_k] \ \parallel\ t' \rightarrow \varnothing \times [t_1, \ldots, t_k, t'] \\ t' \parallel\ \varnothing \times [t_1, ..., t_k] \ \rightarrow \varnothing \times [t', t_1, \ldots, t_k] \end{cases}$$

4 Grammar

A tree is syntactically defined in BNF as follows:

```
tree := value       → t : value [ ]
    | tree tree     → t : tree | tree
    | / tree        → t : '/' | tree
    | tree , tree   → t : tree ‖ tree
    | ( tree )      → t : tree
    ;
```

The right arrow ($\rightarrow$) indicates the tree built for each syntactical construct. The tree whose value is *slash* (*/*) plays a special role in the tree conversion to OSC messages. This role is described in section 6.

5 Values and Evaluation

This section explains how the trees carrying the special *mathematical operators*, *variables* and *expand* special values are evaluated.

5.1 Mathematical Operators

Mathematical operations on trees are seen as operations on their values that preserve the subtrees. These operations include arithmetic and logical operations, trigonometric and hyperbolic functions, exponential and logarithmic functions, power, square root, etc.

We will designate these operations by `op`. Then for 2 trees $t : v \times [t_1, ..., t_k]$ and $t' : v' \times [t'_1, ..., t'_k]$:

$$\texttt{op} \times [t, t'] \rightarrow (\texttt{op}\ v\ v') \times [t_1, \ldots, t_k, t'_1, \ldots, t'_k]$$

5.2 Variables

The special value type *variable* denotes a tree whose value refers to another tree. Evaluation of a variable tree consists in expanding the referred tree at the variable position. Let's define a variable var and a variable tree t' as follows:

$$\begin{cases} var = v \times [t_1, ..., t_k] \\ t' : \$var \times [t'_1, ..., t'_k] \end{cases}$$

then

$$t'_{\{var\}} \rightarrow v \times [t_1, ..., t_k] \mid \varnothing \times [t'_1, ..., t'_k]$$

t'$_{\{var\}}$ denotes the tree t' with an environment containing a definition of the variable var.

Example :

```
x = x 0;
y = y 0;
/A/B $x, $y;    ⇒ /A/B (x 0), (y 1);
```

Local Environnements Each tree is evaluated in an environment containing the list of all the variables of its parent. However, a variable can be evaluated in a local environment, which is defined inside braces:

$$\begin{cases} var = t \\ \$var\{a = t1, b = t2, \dots\} \rightarrow t_{\{a,b,\dots\}} \end{cases}$$

5.3 Expand Value

An *expand value* is a special value that is expanded to a forest. It can also be seen as a *loop* control structure. The syntactic form is as follows:

$id[n \dots m]$ where n and m are integers
$id[ab \dots xy]$ where a, b, x, y are letters.

We will note ε the expansion operation:

$$\begin{cases} \varepsilon(id[n \dots m]) \rightarrow \varnothing\ [id_n, id_{n+1}, \dots, id_m] \\ \varepsilon(id[ab \dots xy]) \rightarrow \varnothing\ [id_{ab}, id_{ac}, \dots, id_{ay}, \\ \qquad\qquad\qquad\qquad \dots, \\ \qquad\qquad\qquad\qquad id_{xb}, id_{xc}, \dots, id_{xy}] \end{cases}$$

where each id_n is a tree $v \times [\,]$ whose value v is the concatenation of the base value id and of the current index n.

Special Forms An *expand value* can also take the following special forms:

$id[i : n \dots m]$ where i is an identifier
$id[i : j : ab \dots xy]$ where i, j are identifiers.

The identifiers denote variables that are instantiated in the environment by the expansion operation, with the current index value. For example:

$$\varepsilon(id[i : n \dots m]) \quad \rightarrow \varnothing\ [id_{n\{i=0\}}, id_{n+1\{i=1\}}, \dots, id_{m\{i=m-n\}}]$$

6 Conversion to OSC

An OSC message is made of an OSC address (similar to an Unix path) followed by a list of data (which can possibly be empty) The *slash* special value of a tree is used to discriminate the OSC address and the data. In order to do so, we type the values and we define @ as the type of a value part of an OSC address.We'll note $type(v)$ to refer to the type of the value v.

We'll note t^a a tree t which value is of type @ . Then we define a @ operation that transforms a tree in to a *typed* tree:

$$@(v \times [t_1, ..., t_k]) \rightarrow \begin{cases} \varnothing \times [t_1^a, ..., t_k^a], v = / \\ v\ \times [t_1, ..., t_k], v \neq / \end{cases}$$

The conversion of a tree t into OSC messages transforms the typed tree $@(t)$ into a forest of OSC addresses followed by data:

$$\texttt{OSC}(v \times [t_1, ..., t_k]) \rightarrow \begin{cases} \varnothing \times [v \times \texttt{OSC}(t_1), \dots, v \times \texttt{OSC}(t_k)], & type(v) = @ \\ v\ \times [\texttt{OSC}(t_1), \dots, \texttt{OSC}(t_k)], & type(v) \neq @ \end{cases}$$

7 Example

The script below presents an example of the new version of the INScore scripting language. Variables are indicated in blue. Local variables are declared in red.

```
# variables declaration
pi    = 3.141592653589793;

# '$step' makes use of 'count' a local variable
step  = / ( * 2, $pi), $count;

# '$i' is defined by the expansion of the address 'n_[i:1...9]'
x = math.sin ( * $step, $i );
y = math.cos ( * $step, $i );

# the following variables select part of guido
# music notation code to build a short score
dyn = (? (% $i, 3), '\i<"p">', '\i<"ff">');
note = (+ $dyn, "␣", (?  (% $i, 2), "e2", "g1/8"));

# this is a classical OSC message that simply clears the scene
/ITL/scene/* del;

# this is the main variable. It will be expanded to create
# a series of small scores. The variables are computed
# using locally defined variables.
notes = (/ITL/scene/$addr
        (set gmn (+ "[", $note, "]")),
        (scale 0.7),
        (x * $x, $radius),
        (y * $y, $radius));

# finally '$notes' is used  with addr, count and radius as local
# variables, which could be viewed as a function call.
$notes{addr=n_[i:1...9], count=9, radius=0.7};
```

Evaluation of this script produces OSC messages fully compatible with the previous version of the language, and which are schematically presented below.

```
/ITL/scene/n_1 set gmn '[ ı<"ff"> g1/8]';
/ITL/scene/n_1 scale 0.7;
/ITL/scene/n_1 x 0.0;
/ITL/scene/n_1 y 0.7;
...
/ITL/scene/n_9 set gmn '[ ı<"ff"> c2]';
/ITL/scene/n_9 scale 0.7;
/ITL/scene/n_9 x -0.411452;
/ITL/scene/n_9 y 0.56631;
```

In practice, this example expresses the score illustrated in Figure 3 in just a few lines.

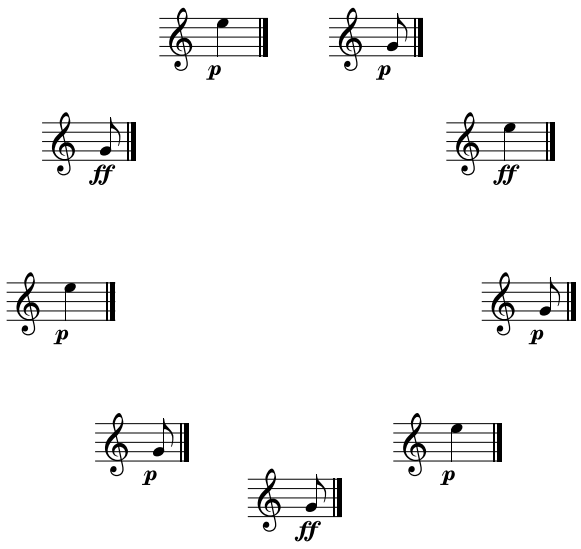


Fig. 3. INScore scene corresponding to the sample script given in section 7.

8 Conclusions

From two elementary operations on trees - sequencing and parallelisation - we have homogeneously introduced the notions of variables and of mathematical and logical operations on trees. The resulting language is much more expressive and more flexible than the previous version of the INScore scripting language. It supports parallelisation of the arguments of a message, variables to describe addresses, series of addresses expressed in a concise manner, use of local variables allowing reusing scripts or parts of scripts in different contexts.

References

1. Antoniadis, P.: Embodied navigation of complex piano notation : rethinking musical interaction from a performer's perspective. Theses, Université de Strasbourg (Jun 2018), https://tel.archives-ouvertes.fr/tel-01861171
2. Daudin, C., Fober, D., Letz, S., Orlarey, Y.: The Guido Engine – A toolbox for music scores rendering. In: LAC (ed.) Proceedings of Linux Audio Conference 2009. pp. 105–111 (2009), lac2009.pdf
3. Fober, D., Letz, S., Orlarey, Y., Bevilacqua, F.: Programming Interactive Music Scores with INScore. In: Proceedings of the Sound and Music Computing conference – SMC'13. pp. 185–190 (2013), fober-smc2013-final.pdf

4. Fober, D., Orlarey, Y., Letz, S.: INScore - An Environment for the Design of Live Music Scores. In: Proceedings of the Linux Audio Conference – LAC 2012. pp. 47–54 (2012)
5. Fober, D., Orlarey, Y., Letz, S.: Scores Level Composition Based on the Guido Music Notation. In: ICMA (ed.) Proceedings of the International Computer Music Conference. pp. 383–386 (2012), icmc12-fober.pdf
6. Fober, D., Orlarey, Y., Letz, S.: INScore Time Model. In: Proceedings of the International Computer Music Conference. pp. 64–68 (2017)
7. Good, M.: MusicXML for Notation and Analysis. In: Hewlett, W.B., Selfridge-Field, E. (eds.) The Virtual Score. pp. 113–124. MIT Press (2001)
8. Hewlett, W.B.: MuseData: Multipurpose Representation. In: E., S.F. (ed.) Beyond MIDI, The handbook of Musical Codes. pp. 402–447. MIT Press (1997)
9. Hoos, H., Hamel, K., Renz, K., Kilian, J.: The GUIDO Music Notation Format - a Novel Approach for Adequately Representing Score-level Music. In: Proceedings of the International Computer Music Conference. pp. 451–454. ICMA (1998)
10. Kuuskankare, M., Laurson, M.: Expressive Notation Package. Computer Music Journal **30**(4), 67–79 (2006)
11. Nienhuys, H.W., Nieuwenhuizen, J.: LilyPond, a system for automated music engraving. In: Proceedings of the XIV Colloquium on Musical Informatics (2003)
12. Roland, P.: The Music Encoding Initiative (MEI). In: MAX2002. Proceedings of the First International Conference on Musical Application using XML. pp. 55–59 (2002), http://xml.coverpages.org/MAX2002-PRoland.pdf
13. Schottstaedt, B.: Common Music Notation., chap. 16. MIT Press (1997)
14. Wright, M.: Open Sound Control 1.0 Specification (2002), http://opensoundcontrol.org/spec-1_0

Surveying Digital Musical Instrument Use Across Diverse Communities of Practice

John Sullivan and Marcelo M. Wanderley

IDMIL, CIRMMT
McGill University, Montreal, Canada
{john.sullivan2, marcelo.wanderley}@mcgill.ca

Abstract. An increasing number of studies have examined the active practice of performers who use digital musical instruments (DMIs) and applied findings towards recommendations for the design of new technologies. However, the communities of practice typically considered in these works tend to be closely aligned with the design communities themselves, predominantly found in academic research and experimental and technology-based music practices. Here we report on an online survey of musicians designed to look beyond these distinct communities to identify trends in DMI use across a wide variety of practices. Compared with current literature in the field, our diversified group of respondents revealed a different set of important qualities and desirable features in the design of new instruments. Importantly, for active and professional performers, practical considerations of durability, portability and ease of use were prioritized. We discuss the role of musical style and performance practice in the uptake and longitudinal use of new instruments, and revisit existing design guidelines to allow for the new findings presented here.

Keywords: digital musical instruments, NIME, survey, communities of practice, design

1 Introduction

The field of digital musical instrument (DMI) design, and much of the music technology domain wherein it resides, can be seen as a dichotomy between multidisciplinary technological research and creative musical practice. This relationship is mutually beneficial, as each side informs the other: innovative technology and design introduce new instruments that augment the capabilities of music production and performance, while expanded musical practice inspires and informs research in new directions. Evaluation of new musical instruments and interfaces is a critical area of research in the field [1], and a focus on embodied, phenomenological perspectives [4, 5] has led to in-depth examinations of communities of practice [7] and the interconnection between performance and design.

The technical definition of a DMI is relatively straightforward, described as an instrument that uses computer-generated sound and consists of a gestural controller to drive musical parameters of a sound synthesizer in real time [8].

In practice, the term DMI, along with the related term "NIME" (when referring to an instrument or interface, ie., *a performer playing a NIME*)[1], is most commonly associated with non-commercial, atypical musical instruments and interfaces that are not generally used, or available, in mainstream music performance.

This constrained scope tends to be transferred to the prevailing research on DMI user groups as well, with most scholarship on DMI performance situated within academic and experimental music contexts. However, beyond these focused communities there is a diverse ecosystem of performers who use instruments that may fit the technical definition of a DMI but not the typical social and cultural context associated with the term.

While studies of DMI-centric musical practice are valuable, they may fail to capture unique and diverse perspectives coming from other communities. For example, electronic dance music (EDM), hip hop, DJs, experimental rock bands and modular synthesizer communities are just a few areas of practice that rely heavily on existing and emerging digital technologies for performance, but are not typically included in the discourse. Input from these groups can broaden the understanding of where and how DMIs are being used in different contexts, and ultimately inform the design and evaluation of new DMIs for successful and long-term use in active musical practice.

To investigate this further we created a new online survey to poll musicians on their use of digital, electric and computer-based instruments in performance. The survey contained a wide range of questions about respondents' backgrounds, performance practices, musical styles, and instruments.

Our work differs from related previous studies by our open invitation for any and all musicians to take part. In the survey we chose to use the term *electronic musical instrument (EMI)* as a generic and inclusive name for various overlapping terminologies used in the field such as DMI, NIME, computer-based instrument, interface, controller, etc. By avoiding domain-specific jargon we hoped to make the survey accessible and applicable to a diverse cross-section of performers.

Here we report on our initial findings from the survey. First, we review related surveys and questionnaires about use of DMIs in performance, which informed the design of our own survey (Sec. 2). Next, we describe the formulation of our questionnaire and how the survey was carried out (Sec. 3). We then share the results of our analysis which was carried out in two rounds (Sec. 4). Finally, we reflect on our findings, comparing them to previous work and reflecting on implications for the design of EMIs intended for long-term use in performance (Sec. 5).

[1] New Interfaces for Musical Expression, coming out of the conference of the same name. (`nime.org`)

2 Related Work

In the interest of providing designers with better tools and more information to aid the creation of new instruments, researchers have utilized questionnaires to survey performers about the use of DMIs in their musical practice. In our own work, we are interested in identifying underlying factors that contribute to the adoption and long-term use – or rejection – of new digital musical instruments. Here we review previous questionnaire-based surveys of DMI performance communities which provided the background for our work. While each had its own specific focus and goals, they all placed a primary focus on the embodied connection between performer and instrument.

2.1 Dual Performer-Designer Roles

In 2006, Magnusson and Hurtado conducted a survey of musicians who play electronic music, with a focus on the differences between acoustic and digital instruments [6]. Respondents to describe the tools they used and the nature of their relationships with them. Participants were recruited through several audio programming mailing lists including the investigators' own audio software mailing list. Accordingly, most respondents were highly computer-literate and skilled computer programmers.

Two particular findings of the survey highlight the specialized nature of the DMI user community that was investigated. First, respondents liked the ability to easily create and modify digital instruments according to specific needs of a performance or composition. The technical knowledge necessary for these "easy" designs and modifications indicate advanced skillsets in various non-musical areas (such as computer science and software development) that are not typical of most musicians. Furthermore, it shows that many of the respondents identify as instrument designers as well as as performers.

Second, the respondents tended to be more critical of digital instruments than their acoustic counterparts. Entropic (non-deterministic) characteristics of digital instruments were generally considered to be flaws or errors in the system, whereas entropy in acoustic instruments was regarded favorably as giving the instrument character leading to discovery of new sounds or playing techniques. This outlook indicates a design-centric evaluation of an instrument, understandable given that most respondents were instrument builders themselves, well-versed in the craft and background research of the field.

In [11] Paine carried out another questionnaire-based study that gathered data about DMIs for the development of a taxonomy for DMIs. As with [6], respondents identified as both performers and designers. Furthermore, they varied in how they thought of or referred to the systems they were discussing: instruments, interfaces, compositions, or something else. The authors observed that the *"...notion of interface/instrument considered also in terms of a composition, while familiar to those working in the area, is of course radically different from the concept of a traditional acoustic instrument."* Again this illustrates how select and idiosyncratic the "typical" DMI performance community is.

2.2 Surveying the NIME Community

A pair of recent surveys elucidate some of the limitations around performance and the continued use of DMIs over time. The first surveyed instrument makers whose instruments had been presented at the NIME conference over several years [9]. This was followed by a survey of NIME performers to explore and understand the roles of DMIs in their practice and understand common values among performers [10]. They confirmed that a majority of new DMIs fail to be developed or used beyond their initial design and infrequent use in actual performance, and identified a few primary factors contributing to this trend: DMIs are often designed as research probes or works-in progress not intended for real-world use; instruments are most frequently used by only one or two performers (and most often the primary/only performer is the designer); instruments frequently suffer from maintenance and reliability issues; perspective performers lack the opportunity to use them in performance.

Common themes that were identified around the use of DMIs included the desire for bespoke instruments that could meet personalized and idiosyncratic needs most commonly associated with performing experimental and exploratory styles of music. Consistent with the other surveys discussed in this section, they also found that most (78%) of the performers who responded had designed their own instrument.

2.3 Beyond NIME

The studies discussed above illustrate an active, engaged, and highly skilled community of performers, researchers and designers. The area has grown and matured, and is a vital contributor to continued innovation in both instrument design and evolving musical practice. However, a vast community of electronic and digital instruments – and the performers that use them – exists outside of these surveyed communities. Whether by virtue of mass appeal and commercial availability, or their use in more conventional and mainstream music communities, perspectives from these populous and highly active communities of digital instrument users are seldom included in DMI user research.

Our investigation in this direction began with a preliminary survey to examine DMI use across widespread communities of practice [13]. A key finding of that work identified the largely pragmatic factors influencing the abandonment new instruments and technologies. This led to our literature-based analysis of essential qualities for DMIs to be viable for use in professional performance situations, most importantly instrument stability, reliability, and compatibility with other instruments, performers and industry standards [14].

3 The Electronic Musical Instrument Survey

Following our previous work, we were interested to conduct a more comprehensive online survey that again targeted performers across a wide variety of

performance practices and focused on factors that contribute to the uptake and continued use of new instruments in performance. Additionally we wanted to compare behaviors and preferences of user groups like those researched in previous works to those operating in more mainstream and popular music circles.

3.1 Participant Criteria and Recruitment

The survey was open to all performers, with no specific requirement that they use electronic musical instruments (EMIs) in performance. The survey was administered online and formatted conditionally so that only those who reported using EMIs saw those relevant sections. Participants were required to be 18 years of age. Beyond that, the only requirement was that respondents identified themselves as "active musicians".

Calls for participation were sent via academic mailing lists and across social media and online music forums to musicians and performance communities. As an incentive for participating, respondents were invited to enter a drawing for a gift certificate to an online music retailer.

3.2 Questionnaire

Our previous survey had used mostly closed and short answer questions to both minimize the length of time to complete the survey (and in doing so, maximize the number of respondents) and to optimize and automate analysis of the data. For this survey we chose to ask more open-ended questions, and conducted qualitative analysis of the free-format responses.[2]

The questionnaire was organized in two parts with a total of four sections. The first section collected demographic (age, gender, location) and background information about the respondents and their musical training, including how long they had been playing music, details on formal training, areas of focus, and experience with computer programming and electronics. Section two asked asked about their performance practice: primary genres and sub-genres of music that they perform, frequency and types of performance, what kinds and sizes of venues, if they play solo or with groups/ensembles, and what kinds of instruments and setups are used.

Part two of the questionnaire was dedicated to the use of electronic musical instruments and controllers. Because the survey was open to all performers, it started with the question, "Do you use electronic musical instruments in performance?" If a respondent answered no, the survey concluded at that point. If they answered yes, they moved to section three, which asked about the types of instruments and controllers they use. They were asked to give information about the instrument or controller they use the most, and could repeat the section up to three times to give information on multiple instruments. Section four of the survey contained several open-ended questions about the respondent's opinions on acquisition and continued use of EMIs.

[2] The questionnaire can be viewed at: `https://emisurvey.johnnyvenom.com/questionnaire.pdf`

3.3 Data Collection and Analysis

A website was built to host the survey and put online at the domain `emisurvey.online`[3]. Responses were saved on a server database, then compiled to a spreadsheet that was downloaded for analysis.

We began our analysis by classifying the participants by background, experience, musical styles, and how active they were as performers. Then we analyzed the respondents' answers qualitatively using techniques taken from Grounded theory [12]. Our methodology used multiple rounds of coding, first open, then using the constant comparison method, where codes between answers and participants were associated into related concepts and themes. This process yielded several high level insights and provided the the motivation and rationale to perform a deeper analysis that focused on the more active performers, those who performed more frequently.

4 Results

A total of 85 people responded (M=60; F=22; other/not specified=3). Respondents were primarily North American and European, and most were between 26 and 65 years old (26 - 45: 65%; over 45: 27%; under 25: 8%). Collectively, the survey population is highly experienced, with 89% reporting more than 10 years of experience in music performance, and 64% more than 20 years. 85% have received formal training with more than a third at or above graduate level.

4.1 Performance Practice

As shown in Table 1, there was a wide range of diversity in the frequency and type of performances across respondents. Over half perform 10 times or less per year. Average audience size varies from less than 100 to over 1000. Most play both solo and in groups.

Performances/year		Avg. audience size		Solo/group performance	
10 or fewer	53%	less than 100	56%	Both solo and group	60%
11 - 20	22%	100 - 500	47%	group only	25%
20 - 50	13%	500 - 1000	16%	solo only	15%
50 or more	12%	more than 1000	8%		

Table 1. Performance frequency, average audience size and configuration of respondents. Multiple answers could be chosen for audience size.

To classify musical styles, we used the list of genres was taken from AllMusic, an online music database[4], with some changes made to reflect some of the tastes

[3] Now archived at `https://emisurvey.johnnyvenom.com/survey-archive/`.
[4] `https://www.allmusic.com/genres`

and nuances of expected respondents. For instance, *electronic* music may mean vastly different things to popular or experimental musicians, so it was divided into *EDM* and *electro-acoustic*. Respondents could choose multiple genres and could specify additional sub-genres or styles. Totals for each category were adjusted to include any sub-genres that we felt belonged in the given categories. The most common styles of music reported were: avant-garde/experimental and electro-acoustic, followed by classical, EDM, rock/pop, jazz, and folk. The full results are shown in Table 2.

Musical Style	Percent	Total	Musical Style	Percent	Total
Avant-garde/Experimental	68%	58	Stage/Theater	8%	7
Electro-Acoustic	34%	29	International	5%	4
Classical	26%	22	Blues	2%	2
EDM	22%	19	Latin	2%	2
Pop/Rock	14%	12	R&B	1%	1
Jazz	12%	10	Rap	1%	1
Folk	11%	9	Country	0%	0

Table 2. Self reported musical performance styles.

The results show significant blending and mixing of genres, especially across and between traditional classifications of "art" music (ie., avant-garde, electro-acoustic) and "popular" music (EDM, rock/pop, etc.) styles [2]. It should also be mentioned that self-categorization of genre and style is extremely subjective, and similar musics may be reported across different categories by different respondents.

75% of respondents use traditional instruments in their performances (played by either themselves or others they perform with). This includes orchestral instruments and typical rock instruments (ie., guitars, drums, etc.), and both acoustic and electric instruments. The full instrument classification is shown in Table 3. Nearly half use computers in performance, and a quarter use DMIs or DIY or self-made instruments. Interestingly, 92% of respondents reported that they have experience with computer programming or electronics.

Our intent was to reach a number of different performance communities, but we still found that many respondents fit into typical DMI-centric performance practices. 68% (58 total) came from formal training and academic settings, were involved in experimental music practices, and were technologically adept. As this study was carried out in an academic research environment, many of the respondents can be recognized as operating in or adjacent to academic practices. Therefore we recognize the implicit bias of our networks through which the survey was distributed, and acknowledge the limits of our attempt to capture a sufficiently broad diversity of performance communities. However, 33% (28) of respondents work across both art and popular music genres, and another 12%

Instrument classification	Percent	Total
Traditional instruments (acoustic and electric)	75%	64
Computers and software	48%	41
Synths/sequencers/samplers and other hardware	35%	30
DMIs and DIY instruments	25%	21
Controllers	21%	18
unspecified electronics	13%	11

Table 3. Types of instruments used in performance.

(10) strictly in popular music genres. Ultimately we found our population significantly diverse, representing a variety of different approaches and perspectives to performance.

4.2 Electronic Musical Instruments

In the second half of the survey, participants were asked if they use electronic musical instruments in performance. Of the 85 total respondents, 23 (27%) answered that they do not, bringing them to the end of the survey. The remaining 62 participants continued to the second half of the survey, where they identified and gave information about their their primary electronic instrument(s) (up to 3), and responded to general questions about instrument uptake and longitudinal use. The instruments were categorized and are shown in Table 4.

Electronic instrument category	Percent	Total
software	71%	44
MIDI controllers	69%	43
keyboard synths	47%	29
FX processors	40%	25
FX pedals	39%	24
samplers	37%	23
drum machines	35%	22
modular synths	31%	19
other	19%	12

Table 4. Primary electronic musical instruments used.

Initial coding of the responses to the remaining survey sections revealed a number of consistent trends across users. Most noticeable was the prevalent use of computer software and MIDI controllers. Asked whether they prefer computers or dedicated hardware for performance, 26% chose hardware and 19% chose

computers, while nearly half said it depends and didn't indicate a preference for one over the other. Positive attributes for hardware included stability and reliability, as well as a preference for tactile controls, imperfections (consistent with [5]), "live-ness and risk-taking", and simplicity of devices used for dedicated tasks. Computers were favored for size, convenience, versatility, affordability compared to the cost of hardware, and ability to handle more complexity than dedicated hardware.

Instrument Satisfaction (and Dissatisfaction) The most common factors that contributed to instrument satisfaction were largely pragmatic: size and portability was the most frequently mentioned, followed by flexibility and versatility, ease of setup and use, responsiveness, and compatibility with other gear and software. Factors that lead to dissatisfaction included a lack of desired features, not enough controls, desire for more flexibility, and desire for better sound quality.

There were differing opinions about flexibility, occasionally from the same participant. On one hand flexibility is desirable for discovery and exploration, as well as plain economics: one versatile piece of gear can do the job of several dedicated devices. On the other, performers appreciate the simplicity and reliability of dedicated devices for specific tasks. Furthermore, dedicated devices may provide useful constraints which can enhance exploration and creativity (as investigated by Zappi and McPherson in [15] and Gurevich, et al. in [3].) One participant pointed out different priorities for composition/production and live performance: flexible instruments are beneficial in the studio but are a liability in live performance, for which they prefer the direct control and reliability of dedicated devices. Interestingly, while many found flexibility to be a desirable quality, most respondents only use basic configuration options that their instruments provide, such as tweaking factory presets and basic parameter mapping.

Uptake, Longevity and Retiring Instruments The most popular reason given for taking up a new instrument was to explore new musical possibilities and expand creative expression. Other frequent reasons were to meet an established compositional or performance goal, to acquire new functionality (new features, workflows or remove restrictions), and to upgrade older gear.

Most respondents reported that there is no time limit on retiring an instrument. If it works and fits within their setup, they will use it until it is no longer functional. Participants cited obsolescence, lack of continued manufacturer support and loss of compatibility as factors that lead to instrument retirement. Another important factor mentioned was evolving musical styles and practices, along with diminishing interest and enthusiasm for an individual's existing instrument, with one participant saying that "new instruments inspire new music."

Two respondents who reported designing their own instruments (or instruments for others) also stated that their instruments are frequently redesigned or in a continual state of development. This behavior is consistent with previous

research within DMI communities (as discussed in Sec. 2) but was uncommon in our results.

4.3 Demands of Performance

Throughout the initial analysis, we noticed distinct differences between the answers of respondents who performed frequently and those who didn't. We then conducted a second round of analysis with only the more active performers. From the 62 respondents who use EMIs, 32 who reported playing 10 or fewer shows per year were removed, leaving 30 "active" performers for analysis. Of the 30, 20 play art music genres, 5 play popular music genres, and 5 play both.

Consistent with the first analysis, these active performers primarily use popular and commercially available hardware and software. None build their own instruments, though two use instruments built for them. Of the non-commercial instruments mentioned, there were three augmented instruments (traditional instruments equipped with sensors to control computer-based audio processing), and one custom built synthesizer. Also consistent with the larger group, the active performers primarily use computers and software in performance. The most common software and languages mentioned were Ableton Live, Max and Pure Data.

There were some important differences as well. When the less active performers were filtered out, much more attention was given to pragmatic issues of functionality for performance like reliability, portability and ease of setup, and less to creative or musical concerns like expressiveness, achieving virtuosity and novel interaction methods. The most common factors influencing instrument choice for active performers were:

- needing flexibility and versatility
- importance of (small) size and portability
- simplicity and ease of setup and use
- potential for exploration and discovery with new instruments
- evolving musical styles and performance practice dictate choices in equipment
- concerns about instrument failure, build quality and reliability
- coping with compatibility issues, connectivity, support, obsolescence

Preference for Computers Active performers indicated a decisive preference for computers and controllers over dedicated hardware, citing simplicity and portability as their biggest advantages. This highlighted the greater technical proficiency in the active performance group versus the rest. Whereas, in the first analysis, some respondents found computer-based performance setups to be unreliable and preferred hardware, the active group indicated the opposite, citing concerns about hardware failure, build quality and reliability, and relying heavily on computer based setups. Technical competence was also indicated with the active performance group reporting much deeper configuration and customization of their instruments and performance setups than the less active performers.

5 Towards Design for Performance

One of the key differences we have found between our results and those of previous studies is that our participants are much less involved in the design and development of the instruments that they use. Most work with popular, commercially available instruments, controllers and software available off the shelf.

Morreale and McPherson's survey on instruments includes design considerations for instruments intended for long-term use [10]. Our results were consistent many of their key concepts: simplicity of interaction, quick and easy set-up, portability, quality and craftsmanship, use of commonly used, stable technologies, and extending musical possibilities. They also identified the appeal of "signature features" and unique aesthetic qualities. Results from our survey showed that these considerations, while mentioned, were secondary to more practical issues related to performing with reliable, functional instruments such as compatibility and flexibility.

6 Final Remarks

Our survey aimed to associate instrument preference and desirable attributes with differences across various types of practice and musical styles that may be less represented in previous research. In doing so, we hope to uncover latent factors across diverse performance practices that could inform the design process of new instruments intended for use in active performance practices.

Our intent was to target a wide diversity of musical practices and styles to compare and contrast with previous user research that has tended to focus on academic and research-based DMI design communities and is aligned with avant-garde and experimental music styles. Our success in this endeavor was mixed, and many of our respondents fit within these conventional DMI research frameworks. However, several others reported active DMI use in other dedicated performance contexts, most notably in popular music performance, and illustrated significant diversity.

Some of our findings were consistent with previous studies, while we found other aspects of DMI performance that should be added to the conversation. Most importantly, we found that there are differing design priorities between individuals who maintain an active performance schedule as opposed to those who perform less frequently.

Continued analysis will aim to more closely associate these results to distinct communities of practice, within and beyond typical NIME and DMI-centric paradigms such as those examined in [7]. For example, previous studies found DMI users are frequently closely associated with, or active in, the design and research of new instruments, however this trend was not reflected in our own results which prioritized active performers. In these cases it becomes important to disentangle the roles of design and creative practice in order to examine DMI use from a purely performer-oriented perspective.

Our current work is focused on the design and longitudinal evaluation of new instruments for performance. Informed by the results shown here, we are running

co-design workshops with performers to develop new instrument prototypes. Multiple iterations will produce a stable, performance-ready instrument to be evaluated by several participants in real-world conditions over several months.

Acknowledgements The authors would like to thank Drs. Andrew McPherson and Fabio Morreale for their valuable input and discussions during this project.

References

1. Barbosa, J., Malloch, J., Wanderley, M., Huot, S.: What does 'Evaluation' mean for the NIME community? In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 151 – 161. Baton Rouge, LA, USA (2015)
2. Glahn, D.V., Broyles, M.: Art music (2012), `http://www.oxfordmusiconline.com/subscriber/article/grove/music/A2227279`
3. Gurevich, M., Marquez-Borbon, A., Stapleton, P.: Playing with Constraints: Stylistic Variation with a Simple Electronic Instrument. Computer Music Journal 36(1), 23 – 41 (2012)
4. Leman, M.: Embodied Music Cognition and Mediation Technology. The MIT Press, Cambridge, MA, USA (2007)
5. Magnusson, T., Hurtado, E., Magnusson, T.: The Phenomenology of Musical Instruments: A Survey. eContact! 10(4), 6–10 (2008)
6. Magnusson, T., Mendieta, E.H.: The Acoustic, the Digital and the Body : A Survey on Musical Instruments. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 94–99. New York City, NY, USA (2007)
7. Marquez-borbon, A., Stapleton, P.: Fourteen Years of NIME : The Value and Meaning of 'Community' in Interactive Music Research. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 307–312. Baton Rouge, LA, USA (2015)
8. Miranda, E.R., Wanderley, M.: New Digital Musical Instruments: Control and Interaction Beyond The Keyboard. A-R Editions (2006)
9. Morreale, F., McPherson, A.: Design for Longevity: Ongoing Use of Instruments from NIME 2010-14. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 192–197. Copenhagen, Denmark (2017)
10. Morreale, F., Mcpherson, A.P., Wanderley, M.M.: NIME Identity from the Performer's Perspective. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 168–173. Blacksburg, VA, USA (2018)
11. Paine, G.: Towards a Taxonomy of Realtime Interfaces for Electronic Music Performance. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 436–439. Sydney, Australia (2010)
12. Strauss, A., Corbin, J.: Grounded theory methodology. Handbook of qualitative research 17, 273–285 (1994)
13. Sullivan, J.: Interaction and the Art of User-Centered Digital Musical Instrument Design. Masters thesis, University of Maine (2015)
14. Sullivan, J., Wanderley, M.M.: Stability, Reliability, Compatibility: Reviewing 40 Years of DMI Design. In: Proceedings of the 15th Sound and Music Computing Conference. pp. 319–326. Limassol, Cyprus (2018)
15. Zappi, V., Mcpherson, A.P.: Dimensionality and Appropriation in Digital Musical Instrument Design. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 455–460. London, United Kingdom (2014)

Systematising the Field of Electronic Sound Generation

Florian Zwißler[1] and Michael Oehler[2][0000-0003-1034-8601]

[1] Technische Universität Dresden, Musicological Department, 01062 Dresden, Germany
[2] Osnabrück University, Musicological Institute, 49074 Osnabrück, Germany
michael.oehler@uos.de

Abstract. There is a striking disproportion between the omnipresence of electronic sound, both in all art forms and everyday life, and the shortage of terminological tools capable of apprehending this phenomenon in a suitably scientifical way. A method is proposed to develop a refined terminology which in turn will provide new tools for a discourse on electronically produced sounds. The key element is an in-depth survey of several electronic music studios: a large proportion of the innovations and impulses that defined a new type of musical practice originated here, however a vast number of their informational resources have not yet been explored. Besides the detailed discussion of the structure and nature of the studio as the "instrument of electronic music", the focus of this approach will be an examination of the actual working processes of each studio as well as the interdependencies between studios. This relates to the transfer of knowledge and technology as well as possible interconnections between composers, technicians, scientists etc. The information gained will be collated in a database using a newly defined classification, which should elucidate lines of development that aren't immediately obvious from the raw data. A synoptic analysis of the collected data will then serve as a foundation for refining the basic terminology concerned with electronically produced sound.

Keywords: Electronic Music, Electronic Studios, Sound Synthesis

1 Background

There is a striking disproportion between the omnipresence of electronically generated sound in both the art and everyday worlds on the one hand and the obvious lack of universal conceptual tools to adequately communicate about this phenomenon and thus systematically explore it on the other hand. There is, however, no shortage of literature on this subject. Various publications document a general interest in a substantial discourse or thematize the lack and at the same time the necessity of a systematization (e.g. Enders, 2017; Kim & Seifert, 2017; Bense, 2013; Holmes, 2012; Kvifte & Jensenius, 2006). The obvious difficulty to speak precisely about electronically generated sounds appears to have existed as long as these sounds themselves. If one avoids direct analogies to the sound world of the classical instrumentarium, then for the classification of electronic sounds and sound generators at least one was and is often held on to systematization grids that were created for precisely this classical

instrumentarium (e.g. MIMO, 2017; Montagu, 2007; Simon, 1992; Kartomi, 1990; Hornbostel & Sachs, 1914). However, this perspective promises little success, since it attempts to depict a phenomenon with a fixed instrument concept that is characterized precisely by the fluidity of its manifestations. As a pointed example: according to this classical logic, the three statements that a sound was produced a) with a Trautonium, b) with a Minimoog synthesizer, or c) using the Max/MSP software platform on an Apple computer must have a comparable information content with regard to the description of this sound. That this is not the case is probably evident.

In the scientific literature, which deals with this topic in detail and with a decided claim to precision, there are manifold starting points, which, however, usually only examine a very narrowly defined question. For example, the discussion of a genre such as musique concrète or electroacoustic music. It can be observed that precisely these genre concepts are often delimited with the help of the phenomena studied (e.g. Ungeheuer, 2002) or selectively focused on individual phenomena (e.g. Bovermann et al., 2017; Cook, 2017, 2002; Miranda, 2017; Miranda & Biles, 2007; Neukom, 2005). Individual phenomena or historical developments are depicted using selected composers as examples (e.g. Blumröder, 2017), or a systematization is carried out on the basis of an already clearly limited scope of validity (e.g. Miranda, 2012; Puckette, 2007; Dodge & Jerse, 1985; Pfitzmann, 1975; Mathews, 1963). Another approach to systematization is the categorization by timbre, as it took place within the framework of musique concrète (e.g. Schaeffer, 1966; Chion, 1983; Landy, 2007) or the general discussion about the nature of the construct timbre (e.g. Stumpf 1890; Licklider, 1951; Krumhansl, 1989). Multidimensional models, such as those already proposed by Schaeffer (1966), are also currently much discussed again (e.g. Reuter et. al. 2018) but are also only used within a rather limited scope of validity.

While the procedures and limitations in the respective research context are usually well comprehensible and meaningful, many papers mention in concrete terms that a corresponding systematization or the search for generalizable concepts is necessary, but that an implementation of this requirement is still pending (e.g. Enders, 2017; Kim & Seifert, 2017; Bense, 2013; Kvifte & Jensenius, 2006). These include, for example, questions about the instrumental nature of electronic sound generators and the associated investigations into the current manifestations of game interfaces for operating such sound generators (cf. Grossmann, 1995; Harenberg, 2012). Apart from the already mentioned problem of the application of instrumental systematics to electronic sound production, the focus in these contexts is often on the immediate surface of electronic sound production (e.g. Roberts, Wakefield & Wright, 2017); substantial questions on the conceptual foundations of the actual sound production of these "instruments", on the other hand, remain mostly unanswered.

2 Aims

This theoretical paper proposes a method for systematising the field of electronic sound generation and sound generators that goes beyond the approaches described above and allows new terminological findings to be derived from it. To this end, the

development of technologies and concepts of electronic sound generation will be investigated from a musicological, practical and compositional perspective. The role of different studios (for electronic music) is of particular importance here, since a large part of the innovations pointing the way for music practice originated from them, e.g. the use of additive synthesis within a compositional framework or the methods of manipulating recorded sound, to name two very early instances. At the same time, numerous source materials available in the studios have only partially been explored, if at all. The implementation of this goal will be realised in a current research project. This article presents the planned method.

3 Method

The planned method is divided into three larger, partly interwoven sub-areas:

Step 1: Research visit and comparative study of exemplary studios (for electronic music). By examining the sources and materials as well as the actual working processes in the respective studios, the hypothesis will be tested to what extent the studios (at different times) can be regarded as a very place for electroacoustic sound generation and thus as an "instrument of electronic sound" itself. Within such an examination, the working processes themselves become a component of the systematics and should therefore be explicitly included in the terminological redefinition (using actor-network-theory; Latour, 2005). The identification of interdependencies between the studios should reveal a possible transfer of technology and knowledge, as well as personal interdependencies of scientists, technicians, composers, etc.

Step 2: Merging the data. The first step here will see the development of a preliminary nomenclature for the classification of data collected in step 1. The data will be stored (continuously) from an early point on in a specially designed database system. In addition to archiving and making the data available to other research groups, the main aim is to identify developments and interrelationships that might otherwise not be immediately recognisable in the data.

Step 3: Synoptic analysis. By means of a synoptic analysis of the collected data and using the database system from step 2, a redefinition of the basic terminology is to take place.

3.1 Studio Survey – Comparative Study on Exemplary Studios

The studio as an acoustic as well as musical production cell can be regarded as the very place of electroacoustic sound generation. A look at the diverse developments that have taken place here over the last one and a half centuries in their various forms of existence shows that a studio could mean something quite different at different points during this period. From the optimized location to the mechanical recording of sound events, the origin of the first radio broadcasts, the experimental home of pro-

gressive composers, the laboratory of many varieties of popular music, to the omnipotent digital production toolbox contained in private laptops: the production conditions are undoubtedly subject to a drastic change and thus also the context in which and from which electronically generated sound is heard and generally perceived.

An overview of the development stages of the studio will therefore serve as an initial interpretation key in order to systematically investigate the conditions of electronic sound in history and today. In detail, it will be necessary to examine in particular those studios whose main focus was and is the production of electronic music. Their connection with compositional concepts and thus with central theoretical considerations in 20th century music history will have to be examined. On the other hand, the role of these studios in interaction with technological developments since 1950 is an orientation point of the investigations, since many of their achievements have in turn been a stimulus for new developments. An example of this are studios in which the pioneering work for the use of computer technology in music was carried out long before its breakthrough on the mass markets. In the field of sound synthesis, for example, the concept of frequency modulation synthesis (FM), which was developed by John Chowning (1973) at the Stanford Artificial Intelligence Laboratory and which shortly thereafter caused a revolution in the commercial synthesizer industry, is worth mentioning. On the other hand, we will have to investigate how many studios at other times could only react to such market developments and how these developments ultimately made many of these studios superfluous.

In order to provide an overview of the working methods, technological equipment and musical alignments of the respective studios, the tools, i.e., all relevant technical equipment and the processes implemented with them, are to be recorded in detail over the entire existence of the respective studio. On the other hand, all persons involved in these developments and their applications are recorded - this includes all relevant figures from composers and musicians to technicians, scientists and researchers. The collected data will provide a substantial insight into the specific interaction of technological and musical-artistic criteria. Since there is no similarly oriented overall description of this studio culture in current research literature, this approach is intended to close the gap and create the basis for all further steps.

A comparable project is the International Electronic Music Catalogue by the English composer Hugh Davies from 1967 (Davies, 1967). Its aim was to compile all works of electronic music existing at that time. For the purpose of classifying the works, he assigned them to the respective studios in which they were created, but the examination or categorization of the studios remains comparatively superficial. In the 1990s there was a similar project at the electronic studio of the TU Berlin: Over a period of eight years Folkmar Hein and Thomas Seelig created the "International Documentation of Electronic Music" (Hein & Seelig, 1996). Similar to Davies' catalogue, the Berlin project was also conceived as a reference work in terms of its fundamental focus. Moreover, both studies remain oriented towards the works and do not focus on the applied sound production methods. The present project, on the other hand, is intended to deal precisely with these.

The method of a detailed investigation outlined here makes it indispensable to make a selection from the large number of production sites. The duration of their

existence serves as a decisive criterion: studios that have been active for several decades or are still active are to be preferred as representative data sources to those with only a short lifespan. Another criterion is the total number of works and projects realized there. This is based on the comprehensive (work-centred) study by Hein and Seelig (1996), which was recently published online in an updated version (as of October 2018) (www.emdoku.de). The selected studios are or were owned by radio stations, universities, foundations or commercial operators. Small and private studios are excluded from the selection. This is not least because internal and external communication is supposed to be a decisive object of investigation. Finally, it is ensured in advance that a data situation suitable for the investigation can be found on site.

3.2 Development of a Provisional Classification System

A Basic Catalogue of Questions. In order to classify the data collected during the studio visits, a preliminary systematisation is to be drafted. For example, the equipment will be classified according to its concrete technical functionality (sound generation vs. sound processing vs. sound storage etc.). In a next step, the possibilities of the specific use of such devices in the network are to be recorded. One example would be the process of additive sound synthesis, as it was realized in the 1950s by continuously repeated copying processes using a sine wave generator, a mixer and tape recorders. The adjustment and classification of possible sound generation processes under the aspect of relevant timbre systematics will also be carried out at this point.

Such an approach, which thus goes beyond a mere inventory from the outset, could begin with model-like questions such as the following:

How is the technical apparatus organized?
How are developments of specific control structures driven forward?
From when and at what points are computers used for this purpose?
In which steps do processes of digitization take place?
To what extent are these developments linked to specific concepts of sound generation or to compositional and aesthetic considerations?

Between the collection of the data according to a systematization and the creation of this systematization by the collected data, interdependencies will arise that have to be carefully balanced. It is to be expected that new insights will emerge after the first studio stays, which will necessitate a re-evaluation of what has already been worked out. Data previously considered irrelevant or a revision of the systematization grids could mean that individual work steps have to be modified from the ground up.

Creation of a Database System. A database system is to be set up as a supporting tool for merging the data. On the one hand, this serves to provide the results in the sense of research data management. On the other hand, the collected historical information on the studios (e.g. technologies used, timbre parameters, composers/scientists who have worked in the studio or details on the scientific and artistic output) as well as relevant media files such as photos, videos, sound recordings, scores or program

code are to be structured in this system. Such a (later) publicly accessible system is not only important for the purpose of archiving and providing data for other research groups, but also an important tool for gaining scientific knowledge from the unpublished sources that are scattered and often only exist in the individual studios.

In an iterative process, the system is adapted to the continuously generated data over the entire project period. The DFG project Harmonic Structures [MU 2686/7-1, KL 864/4-1] (2014 - 2018) is a comparable example of how innovative concepts of digital data structuring in the field of musicology make new findings possible.

3.3 Synoptic Analysis and Redefinition of Basic Concepts

Evaluation of the Data. The synoptic overview of the collected data can open up some promising perspectives of comparison: A look at the technical developments within each studio from a diachronic perspective will provide initial insights into the characteristic developments of the individual production sites. The next step will connect this information with the persons involved, their artistic and technical backgrounds and their conceptual and aesthetic decision criteria. Furthermore, special attention will be paid to the links between these decisions and other thematic complexes within the respective temporal contexts: Where and through whom are impulses given for technical innovations? Conversely, how are working methods transformed by technical innovations? How relevant are these for the development of other genres and fields of action?

From the latter question, interactions with other studio cultures, primarily of course from the various genres of popular music, as well as with the steadily growing sector of commercial music electronics over the course of the later 20th century, can be examined. It can already be foreseen that a rich as yet unforeseeable field of further questions can be expected here. An exhaustive discussion of this terrain would certainly take the present concept beyond the limits of its capacities, but it creates a perspective for further considerations and subsequent research projects. Using newly developed search and evaluation functions within the data collection which are adapted to the specific data, links are to be identified which, among other things, help to answer the prototypical questions formulated above.

Terminological Considerations. The last step will be dedicated to the analysis of existing terminological imprecisions and the expansion of the existing vocabulary. The fundamental resource is the combination of different perspectives achieved in the preceding work steps; namely the historical view of the development steps of individual studios as well as the consideration of the interdependencies between the work of these studios and the circumstances surrounding them. In this way, problems of systematization, terminological blurriness within and across genres and the specification of technologically and aesthetically oriented questions can be addressed. A possible result of such a process is an updated classification logic of electronic sound generation. The methods described at the beginning of Section 3.3 form the methodological framework for the processing.

4. Conclusions

The obvious gaps in the terminology of electronically generated sounds demonstrate the need for fundamental research and expanded discourse in this field. The presented approach is designed to address this issue and to work out methods of electronic sound generation and their embedding in historical and current discourses. These methods will be compiled centrally in the light of their artistic and technical framework and systematised in a database system. The creation of such a data basis will be an approach on the way to a redefinition and refinement of terminological tools. Furthermore, the publication of the database will create new opportunities for contributions from a wide variety of research areas.

Acknowledgements

Financial support by the German Research Foundation through grant 42473404.

References

Bense, A.T. (2013). Musik und Virtualität. Digitale Virtualität im Kontext computerbasierter Musikproduktion. Osnabrück: epOs-Verlag.

Blumröder, C.v. (2017). Die elektroakustische Musik. Eine kompositorische Revolution und ihre Folgen. Wien: Verlag Der Apfel.

Bovermann, T., de Campo, A., Egermann, H., Hardjowirogo, S. I., & Weinzierl, S. (Eds.). (2017). Musical Instruments in the 21st Century: Identities, Configurations, Practices. Berlin: Springer.

Chion, M. (1983). Guide des Objets Sonores: Pierre Schaeffer et la Recherche Musicale. Paris: Buchet/Castel.

Chowning, J. M. (1973). The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. Journal of the Audio Engineering Society, 21(7), 526-534.

Cook, P. R. (2002). Real Sound Synthesis for Interactive Applications. CRC Press.

Cook, P. R. (2017). 2001: Principles for Designing Computer Music Controllers. In A. R. Jensenius, & M. J. Lyons (Eds.). A NIME Reader (S. 1-13). Springer.

Davies, H. (Ed.). (1967). Répertoire International des Musiques Electroacoustiques: Int. Electronic Music Catalog. Groupe de Recherches Musicales de l'ORTF.

Dodge, C., & Jerse, T. A. (1985). Computer Music: Synthesis, Composition, and Performance. New York: Schirmer.

Enders, B. (2017). From Idiophone to Touchpad. The Technological Development to the Virtual Musical Instrument. In T. Bovermann et al. (Ed.). Musical Instruments in the 21st Century (S. 45-58). Berlin: Springer.

Großmann, R. (1995). Sechs Thesen zu musikalischen Interfaces. Interface, 2, 155–162. Hamburg

Harenberg, M. (2012). Virtuelle Instrumente im akustischen Cyberspace. Zur musikalischen Ästhetik des digitalen Zeitalters. Bielefeld: transcript

Hein, F., & Seelig, T. (1996). Internationale Dokumentation Elektroakustischer Musik. Friedberg: Pfau.

Holmes, T. (2012). Electronic and Experimental Music: Technology, Music, and Culture. Abingdon: Routledge.
Hornbostel, E. M. von, & Sachs, C. (1914). Systematik der Musikinstrumente. Ein Versuch. Zeitschrift für Ethnologie, 46, 553–590.
Kartomi, M. J. (1990). On Concepts and Classifications of Musical Instruments. Chicago: The University of Chicago Press.
Kim, J. H., & Seifert, U. (2017). Interactivity of Digital Musical Instruments: Implications of Classifying Musical Instruments on Basic Music Research. In T. Bovermann et al. (Eds.). Musical Instruments in the 21st Century (S. 79-94). Springer.
Krumhansl, C. L. (1989). Why is Musical Timbre so Hard to Understand. Structure and Perception of Electroacoustic Sound and Music, 9, 43-53.
Kvifte, T., & Jensenius, A. R. (2006). Towards a Coherent Terminology and Model of Instrument Description and Design. In Proceedings of the 2006 Conference on New Interfaces for Musical Expression (S. 220-225). IRCAM—Centre Pompidou.
Landy, L. (2007). Understanding the Art of Sound Organization. MIT Press.
Latour, B. (2005). Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press.
Licklider, J. C. R. (1951). Basic Correlates of the Auditory Stimulus. In S. S. Stevens (Eds.). Handbook of Experimental Physiology (S. 985–1039). New York.
Mathews, M. (1963). The Digital Computer as a Musical Instrument. Science, 142(3592), 553–557.
MIMO – Musical Instrument Museums Online, http://www.mimo-db.eu, retr. 4.5.19.
Miranda, E. R., (2017). Guide to Unconventional Computing for Music. Springer.
Miranda, E. R. (2012). Computer Sound Design: synthesis techniques and programming. Oxford: Taylor & Francis.
Miranda, E. R., & Biles J.A. (Eds.). (2007). Evolutionary Computer Music. Springer.
Montagu, J. (2007). Origins and Development of Musical Instruments. Lanham, Maryland: Scarecrow Press.
Neukom, M. (2005). Signale, Systeme und Klangsynthese: Grundlagen der Computermusik. Frankfurt: Peter Lang.
Pfitzmann, M. (1975). Elektronische Musik. Stuttgart: Telekosmosverlag.
Puckette, M. (2007). The Theory and Technique of Electronic Music. London: World Scientific Publishing.
Reuter, C., Czedik-Eysenberg, I., Siddiq, S., & Oehler, M. (2018). The Closer the Better: The Role of Formant Positions in Timbre Similarity Perception and Timbre Classification. TIMBRE 2018 McGill Music Conference, Montreal, Canada.
Roberts, C., Wakefield, G., & Wright, M. (2017). 2013: The Web Browser as Synthesizer and Interface. In A NIME Reader (S. 433-450). London: Springer.
Schaeffer, P. (1966). *Traité des Objets Musicaux, Essai Interdisciplines*. Le Seuil.
Simon, P. (1992). Die Hornbostel/Sachs'sche Systematik und ihre Logik. *Instrumentenbau-Zeitschrift—Musik International, 46*(7–8), 64–66.
Stumpf, C. 1890. *Tonpsychologie*. Leipzig: Hirzel.
Ungeheuer, E. (Ed.). (2002). Handbuch der Musik im 20. Jahrhundert (Band 5). Elektroakustische Musik. Laaber-Verlag.

Movement Patterns in the Harmonic Walk Interactive Environment

Marcella Mandanici[1] and Cumhur Erkut, Razvan Paisa, Stefania Serafin[2]

[1] Conservatory of Music "Luca Marenzio", Dept. of Music Education, p.zza A.B. Michelangeli 1, 25121 Brescia, Italy
[2] Multisensory Experience Laboratory, Aalborg University Copenhagen A.C. Meyers Vænge 15, 2450 Copenhagen SV, Denmark
mmandanici@gmail.com

Abstract. "Harmonic Walk" is a responsive environment where users accompany chords to a given tonal melody by moving to specific points in the environment. To help users in this melody harmonization task, this paper suggests two ways that rely on spatial sound or on motor resonance. An empirical comparison between these two cues is the main focus and contribution of the paper. Results indicate that sound spatialization causes no difference in task performance, whereas the movement patterns provided to the participants show significant difference in task performance. These findings seem to indicate that movement pattern is the most efficient way to communicate information and to foster learning processes in the "Harmonic Walk" environment.

Keywords: Full-body interaction; melody accompaniment; spatial audio; motor resonance.

1 Introduction

Harmonic progressions are driven by a sense of expectation deriving from the sound characteristics of a particular musical chord [1]. In the case of the perfect cadence, the tension generated by the dominant chord implies the reaching of a goal (the tonic), and this is perhaps the most common instance of directed motion in tonal harmony [2].

This paper introduces the "Harmonic Walk", a large-scale responsive environment for learning and practicing the tonal harmony and melody accompaniment. In the environment, harmonic changes are triggered at specific interactive points in the physical space, allowing the users to build a perceptual map of the musical chords on the responsive floor. Based on this map, users walk form one point to the other to "play" the harmonic changes. This paper suggests two ways to help users in melody harmonization task: the first relies on spatial sound, and the second is based on the concept of *motor resonance* [3]. While spatial sound cues imply that a user decides when and where to move by using the auditory modality, the motor resonance concept suggests a cross-modal interaction between the

auditory system, visual movement observation, and the motor response. An empirical comparison between these cues during a simple harmonization task is the main focus and contribution of this paper.

The structure of the paper is as follows. Section 2 provides background on the "Harmonic Walk" environment, as well as motion redirection through spatial audio and by motor resonance. Section 3 describes the technical implementation of the large-scale responsive system with motion tracking and spatial audio components. Section 4 describes the empirical tests, namely a between-subjects experiment (exp. 1) and a within-subjects experiment (exp. 2). The results are presented and discussed in Section 5. Finally, implications for design and teaching practices conclude the paper.

2 Background

"Harmonic Walk" is a large-scale responsive environment supported by a motion tracking system. Six interactive landmarks are positioned on the floor surface, each corresponding to numbered musical chords and to the harmonic functions depicted in Figure 1. The six chords represent the harmonic space of a major

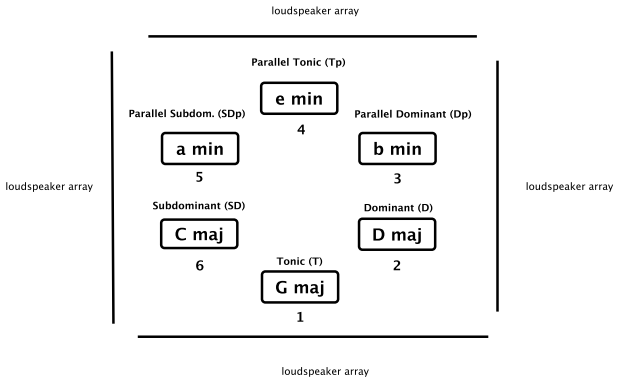


Fig. 1: The "Harmonic Walk"'s interactive space map, with the four loudspeaker arrays for wave field synthesis and the six interactive landmarks for a G major melody harmonization.

tonality and can fully accommodate the harmonization of a melody written in the same key. The users of "Harmonic Walk" have to step from an interactive landmark to another to perform the melody harmonization, exactly in the same way as guitar players move their left hand position on the fingerboard to change the musical chords. Thus, to accomplish the melody harmonization task, users need to know when and where to move to trigger the right chord on the interactive area [4]. This is a complex task whose success depends on the perception of

appropriate metrical and harmonic frames that allow the listeners to group the melodic elements under different harmonic unities, which determine the chord changes [5]. The points where the musical chords change are marked with the corresponding syllables of the song that will be called from now on the "syllables of change" (see Figure 2 for an example). Assuming that the musical features of the melody are acquired through the previous knowledge of the song - or through repeated exposures to it - and that the responsive environment is widely explored by the users, authors concentrate on the processes that govern the acquisition of the motor schemas necessary to harmonize the melody.

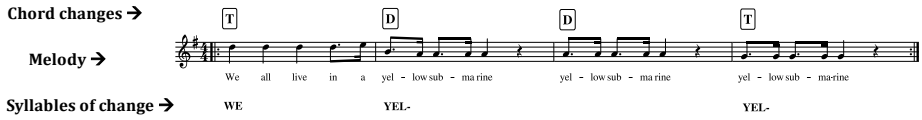


Fig. 2: An example of tonal melody (the chorus of the Beatles' song "Yellow Submarine"), with the chord changes and the syllables of change. The chords employed are the Tonic (T, position 1) and the Dominant (D, position 2).

2.1 Motion Redirection through Spatial Audio

According to Gaver [6] auditory information contains many meaningful data about sound events such as sound source, environmental imprinting and source localization. In real life these data may attract users' attention and produce changes in thoughts, feelings, memory and behavior [7]. The integration of 3D audio and visual elements is crucial to foster a sense of presence and to facilitate users navigation and wayfinding in virtual and augmented reality environments [8]. Particularly, spatial audio has been proven to be more efficient than unlocalisable audio for attention redirection in real and virtual environments [9]. In the "Harmonic Walk" environment authors exploit this property to influence users movement with the aim to help them to locate the right interactive landmark for melody harmonization. The relationship between spatial audio and motion redirection is widely studied for guiding blind people navigation [10] or to build gameful environments to train children to avoid veering [11]. The capacity of sound to influence human movement has also been observed in dancers' reactions in an interactive environment where the sound is produced according to dancers' movements [12]. Dancers build a bodily knowledge of the environment and employ a cognitive map of sounds locations to decide their preferred sound production.

2.2 Learning Interaction through Motor Resonance

Grounded on the discovery of the mirror neurons, the motor resonance theory defines modalities and conditions of imitative behavior. This happens when there is a similarity between a perceived act and the act that one performs [13]. Motor resonance has an important role in social interaction and in sharing intentions and feelings [14]. As movement observation and execution share the same neuronal mechanisms, motor resonance is also a facilitator in sensorimotor learning [3]. This feature seems of particular interest in a large-scale interactive environment such as "Harmonic Walk", where a meaningful interaction depends on the learning of complex sensorimotor schemas.

3 The Harmonic Walk Interactive Environment

The system used for tracking, position detection and audio playback involves two computers running OS specific applications (Figure 3), connected to the same Local Area Network (LAN). Position-based tasks are handled by a Windows PC, while the spatial audio reproduction is executed by a Mac machine running WFS Collider[3].

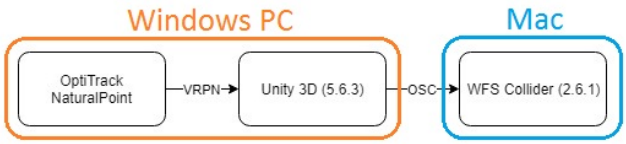


Fig. 3: "Harmonic Walk" interactive environment block diagram.

3.1 Position Tracking

Participant's position is tracked using a system composed of an array of 16 OptiTrack Flex 3 100FPS infrared (IR) cameras arranged in a rectangle placed 2.3 meters above the ground. The cameras were calibrated to track the entire listening space of 2.3*2.3*2.3 meters, with high accuracy (tracking error <2mm). The system sampled the space at a constant 95+ frames/second, with a varying latency of 7-10ms. The system requires one fixed geometry IR reflective marker to be predefined as a trackable item, using the Optitrack NaturalPoint software[4]. This trackable was placed on a hat that participants had to wear. In order to have an interactive system, the first step was to broadcast the position captured

[3] https://sourceforge.net/projects/wfscollider/
[4] `https://optitrack.com/software/`

by the OptiTrack system to Unity3D game engine[5] via VRPN protocol[6]. Inside Unity3D, a scene replicating the physical dimensions of the listening space was created and populated with spherical 3D objects that represent the six chords listening positions (Figure 1). Communication between Unity3D and WFSCollider was established over LAN, using OSC[7].

3.2 Spatial Audio

Spatial audio was reproduced by using wavefield synthesis (WFS), a method used to recreate an accurate replica of a sound field using the theory of waves and generation of wave fronts [15]. WFS affords the reconstruction of sound fields with natural temporal and spatial properties within a volume or area bounded by arrays of loudspeakers [16]. The employed hardware setup included four horizontal arrays of 15 M-Audio BX5 D2 studio monitors + one "the box pro Achat 108 Sub A" subwoofer, digitally summing the signals from each array using the RME TotalMix software. The digital-to-analog conversion was handled by a RME MADIface USB interface, controlling two D.O.TEC ANDIAMO 2 Digital-to-Analog converters (DAC). The software responsible for wavefield synthesis is called WFSCollider, an adaptation of the popular audio programming language SuperCollider. It was developed by Arthur Sauer and Wouter Snoei at The Game of Life Foundation [8] and it is customized for WFS and other object based spatial audio techniques. WFSCollider is inspired by the Digital Audio Workstation workflow, with a timeline, multi-track setup, busses and signal chain racks.

4 Assessment

To accomplish the melody harmonization task in the "Harmonic Walk" environment, users need to acquire a sensorimotor movement pattern related to full-body movements in the responsive space. These movements must fit the musical characteristics of a melody, which determines where and when to move. To help users in making decisions about a convenient movement pattern, authors adopt two different means: one is spatial audio and the other is motor resonance. Two different experiments were organized at the Multisensory Laboratory (MEL) at Aalborg University (Copenhagen, Denmark) with the aim of verifying the effects of these two different conditions on the behavior of participants.

[5] `https://unity3d.com/`

[6] The Virtual-Reality Peripheral Network (VRPN) is a network-based interface for accessing virtual reality peripherals in VR applications (`https://github.com/vrpn/vrpn/wiki`)

[7] Open Sound Control (OSC) is a protocol for communication among computers, sound synthesizers, and other multimedia devices. Further references can be found at `http://opensoundcontrol.org/introduction-osc`

[8] https://github.com/GameOfLife/WFSCollider

4.1 Material

For both experiments authors employed the chorus melody of "Yellow Submarine", a famous Beatles song written in 1966. Thanks to its popularity, only 1 participant over 31 (3.22% of the sample for both experiments) declared not to know it. As can be seen from Figure 2 the melody is a refrain of 4 repeated bars. In the original version only two musical chords are employed for its harmonization, the Tonic (T) and the Dominant (D) whose changes occur on the syllables (or words) WE, YEL. As the Tonic holds number 1 and the Dominant number 2, to harmonize this song it is necessary to find, beyond the position of the Tonic (the starting point), the position of the Dominant (at the top right of the starting point). The movement schema from the Tonic to the Dominant and viceversa is depicted in Figure 4, where the black headed arrows show the movement T-D and the white headed ones the inverse (D-T). The trajectories of the melody

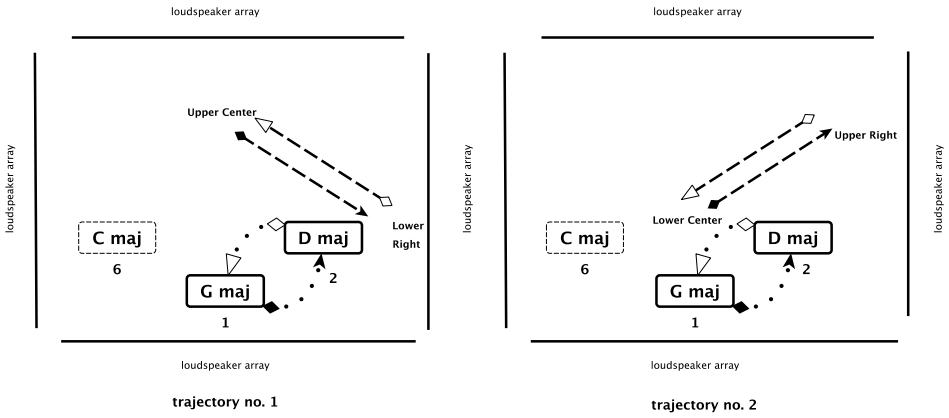


Fig. 4: Two reproductions of the spatial disposition of the Tonic and Dominant chord in the "Harmonic Walk" environment with two different melody source trajectories used to suggest to the users the position of the Dominant chord.

source were designed with the aim of suggesting to the user the idea of reaching the position of the dominant chord. We employed two different approaches to solve this problem: In the first one only the absolute position of the dominant chord is considered. The melody source moves from the upper center towards the position of the dominant chord not following the movements of the user but coming from the opposite direction. Doing so, when moving back, the source must necessarily roll away from the user. The second one is bodycentric, as the melody source follows the relative movements of the user. The source moves according to the relative positions of the chord, from a lower-central position to the upper right and viceversa.

4.2 Validation of the Spatialization Model

To decide what trajectory better conveys the idea to go towards the right, we performed a test with 18 subjects (5 females, age 19-35, $M = 22.5$) chosen from students and staff of Aalborg University. Each subject listened to 4 trials. In each trial the "Yellow Submarine" chorus was played 2 times with trajectory 1 and two times with trajectory 2 in random order. After each trial subjects declared what spatial model better fits the idea of going to the right. 44 answers were for trajectory 1 and 28 for trajectory 2. The chi-square test of independence shows that the differences are significant χ^2, $(71, N = 72) = 7.11$, $p < .01$. Thus trajectory 1 was used as a spatial model for the test.

4.3 Experimental Setup

Table 1 reports the characteristics of the two experiments organized with the aim of studying the impact of 3D audio and of motor resonance on the performances of the participants. The two experiments had essentially the same task that is harmonizing the chorus of "Yellow Submarine" with "Harmonic Walk". In exp. 2, two different versions of the chorus melody were played: one version with the sound source of the melody moving according to the spatial model validated above, and the other version with the melody played still, placing the source exactly at the middle of the responsive space. From now on these two versions will be called SP (spatialized melody) and NSP (non-spatialized melody). In exp. 1 only the SP version of the melody was employed. Moreover, before exp. 2, participants took part in a 45 minutes lesson where the theoretical background of "Harmonic Walk" was presented. Also three examples of melody harmonization with bodily movements were proposed and practiced in the class. Among these there was also the refrain of "Yellow Submarine", but not the chorus. While these two song excerpts sound very similar, the movement schemas for their harmonization are quite different. The schema of the chorus implies an alternance of the Tonic and the Dominant with the pattern 1-2-1-2-1 which from now on will be called the "Song model". For the refrain the schema implies also the Subdominant (position 6) with this pattern 2-1-6-2-1-6-2 which will be called from now on the "Example model". No demonstration of harmonization pattern was provided to participants of exp. 1. Thus essentially the differences between the NSP and the SP condition may emerge from the comparison of the two groups of exp. 2, whereas the impact of motor resonance can be obtained from the differences between exp. 2 and exp. 1.

4.4 Procedure

After registering their personal data, level of music abilities and other information, participants entered the responsive area and explored the environment in order to test how the system reacts to movement and to hear the sound of the six chords of the harmonic space. At the end of this exploratory phase (no more than 3 minutes long) they listened to the chorus melody played with a piano

	Experiment 1	Experiment 2
Type	Between-subjects	Within-subjects
Participants	18 (5 females, age 23-56, $M = 32$) from students and staff of AAU (Copenhagen)	13 (2 females, age 21-35, $M = 26.7$) from a Music Perception Class Medialogy AAU (Copenhagen)
Music Abilities	10 musicians Years of experience, $M = 13.8$ Level of practice, $M = 1.3$[a]	11 musicians Years of experience, $M = 13.54$ Level of practice, $M = 2.4$[a]
Conditions	The melody was always played only in the SP condition	The melody was always played both in the SP and in the NSP condition
Task	Melody harmonization: 3 trials for each subject	Melody harmonization: 4 trials subject, 2 with SP audio and 2 with NSP audio in random order

a Practice is calculated in this way: 3 points for practising more than 3 times in a week, 2 points for once in a week and 1 point for less.

Table 1: The setups of the two experiments with "Harmonic Walk".

timbre without any accompaniment (max 2 times) and identified the syllables of change with the help of the test conductor. Depending on the experiment, participants tried to harmonize the chorus of "Yellow Submarine" 3 or 4 times. At each trial a timestamp file was created with the recordings of the positions occupied by participants during the test, at the rate of 10 per second.

4.5 Method

Data are analyzed in the following ways:

1. Position analysis: the number of triggered positions of the six interactive landmarks and the number of movements from the tonic to the chords put at its right (positions 2 and 3) in the first harmonic change.
2. Movement sequences analysis: movement patterns are identified and their frequency is calculated.

Position Analysis. For each trial the system outputs a vector containing the sequence of the positions triggered by the participants. By filtering out the repetitions authors obtain a vector with the position sequence of every participant and create a position matrix composed of all vectors for analysis. A first outcome is the frequency of each triggered position, which can account for the global movement distribution in the experimental sample. The "Song model" requires a movement from the Tonic towards the right (Dominant, position 2) and this is the suggestion authors intended to convey through the use of spatial audio. A second outcome of position analysis is the distribution of the first harmonic change from the Tonic towards the left or the right part. As further harmonic

changes can be biased by the direction taken in the first one, authors evaluated the movement of the first harmonic change as the most meaningful to understand the participants intentions.

Movement Sequences Analysis. This analysis is based on the search of pattern sequences derived from the observations of participants behavior during the test. The identified patterns are three: one is the "Song model", which corresponds to the chord sequence that fits the songs melody (1-2-1-2-1). The second is the "Circular model", which corresponds to a sequence obtained by triggering the chord position one after the other following the circular shape of the spatial arrangement (1-2-3, 1-2-3-4, 1-2-3-4-5, 1-2-3-4-5-6). The third is the "Example model" given during the lesson to the participants of exp. 2 (2-1-6-2-1-6). A string and substring search has been done on the position matrix to calculate the frequency of each model.

5 Results

As explained in Section 4.3, we present first the differences between the NSP and the SP condition (exp. 2) to evaluate the impact of spatialized audio and then compare the global results of exp. 2 with those of exp. 1 to evaluate the impact of motor resonance. The number of triggered positions in the six interactive landmarks and of movement patterns in the NSP and SP condition are depicted in Figure 5, which shows nearly the same results in the two conditions. The distribution of the movements from the Tonic towards the right or the left side in the first harmonic change (first two columns of Table 2) shows that the movements to the right part are a little more for SP, but the difference is not statistically significant ($\chi^2(1,\, N = 46) = 1.24,\, p > .05$). The positions for both exp. 2 (global) and 1 are calculated and visualized in Figure 6, while the number of movements from the Tonic to the right part and to the left part in the first harmonic change are visualized in columns 3 and 4 of Table 2. No one of these differences are statistically significant (paired sample t-test, $t(5) = -.61, p > .05$ for the position

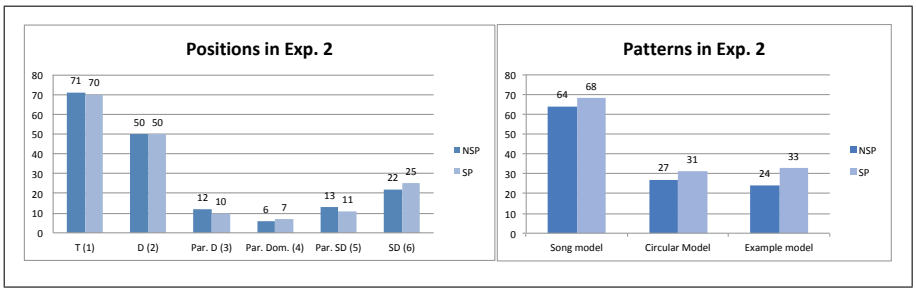


Fig. 5: Results of experiment 2: at the left the triggered positions and at the right the patterns in the NSP and SP condition.

	Experiment 2 (%)		Experiment 2	Experiment 1
	NSP	SP	% (global)	%
R	74	87	80	76
L	26	13	20	16

Table 2: Movements from the tonic to the right side of the environment (positions 2 and 3) and to the left side (positions 6 and 5) in the first harmonic change for the two conditions of exp. 2 and for exp. 2 (global) and 1.

differences and chi-square test of independence, $\chi^2(45, N = 46) = .074, p > .05$ for the movements to the right part).

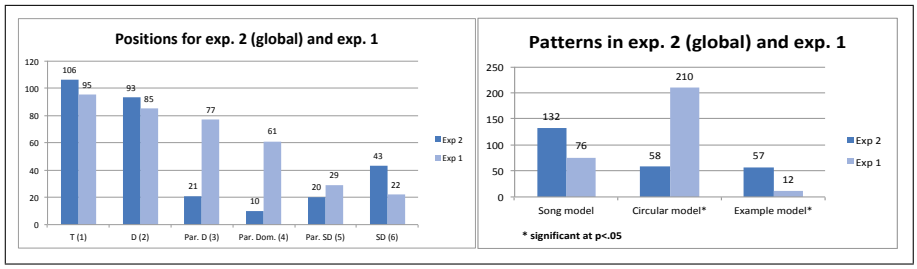


Fig. 6: The results of exp. 2 (both conditions) and 1 for the number of triggered positions and of movement patterns.

The paired-samples t-test between the patterns of exp. 2 and exp. 1 indicates that differences are not statistically significant for the "Song model" ($t(4) = .85, p > .05$). Instead the differences are significant for the "Circular model" ($t(14) = -2.59, p < .05, d = 1.01$) and for the "Example model" ($t(5) = 2.51, p < .05, d = 1.77$)[9].

5.1 Analysis of Results

Spatial audio did not help participants of exp. 2 in harmonizing the chorus of "Yellow Submarine", since the differences in the "Song model" are not significant between exp. 2 and exp. 1. The data of the "Circular model" show that this is much more widespread in exp. 1 that in exp. 2 (Figure 6). This can also be observed in the chart of the triggered positions which are more uniformly distributed in exp. 1 than in exp. 2.

[9] The value of d refers to Cohen's effect size which can be interpreted as (0.2) small, (0.5) medium, (0.8) large.

This behavior may depend on the bias of the circular shape of the interactive landmarks on the application's floor. If participants have no idea of how to move, they simply follow the circle of the interactive landmarks [4] and this is the case of a great number of exp. 1 participants who were not given any movement example. On the contrary, participants of exp. 2 where given a movement pattern (the "Example model") which was used by a lot of them also if this is the motion pattern of the refrain of "Yellow Submarine" and not the model of the chorus (the "Song model"). Thus, the analysis of the "Example model" seems to confirm the influence of motor resonance.

6 Conclusion

This paper described melody harmonization experiments conducted in the responsive "Harmonic Walk" environment. The participants heard a simple monophonic melody, and were asked to harmonize this melody by walking in the responsive environment. Possible musical chords were marked on the floor of the environment, and the participants heard the corresponding chord when they have arrived to a mark from the previous one. Their walking was directed by spatial audio and/or by motor resonance based on previously acquired motion patterns. While spatial audio resulted to have no effect on participants performance, motor resonance showed a great effect in influencing participants behavior.

If confirmed by further studies, this can represent a key point for instructional strategies and didactic action in technological environments such as the "Harmonic Walk". Incorporating motion patterns efficiently in pedagogical instruction will be the first direction in future investigation.

References

1. Cohen, D. E. (2001). The Imperfect Seeks Its Perfection: Harmonic Progression, Directed Motion, and Aristotelian Physics. Music Theory Spectrum, 23(2), 139-169.
2. Rosner, B. S., & Narmour, E. (1992). Harmonic closure: Music theory and perception. Music Perception: An interdisciplinary Journal, 9(4), 383-411.
3. Mènoret, M., Curie, A., des Portes, V., Nazir, T. A., & Paulignan, Y. (2013). Motor resonance facilitates movement execution: an ERP and kinematic study. Frontiers in human neuroscience, 7, 646. doi:10.3389/fnhum.2013.00646
4. Mandanici, M., Rodà, A., & Canazza, S. (2017). Bodily Interactions in Motion-Based Music Applications. Human Technology, 13.
5. Povel, D. J., & Jansen, E. (2002). Harmonic factors in the perception of tonal melodies. Music Perception: An Interdisciplinary Journal, 20(1), 51-85.
6. Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. Ecological psychology, 5(1), 1-29.
7. Truax, B. (1996). Soundscape, acoustic communication and environmental sound composition. Contemporary Music Review, 15(1-2), 49-65.
8. Gunther, R., Kazman, R., & MacGregor, C. (2004). Using 3D sound as a navigational aid in virtual environments. Behaviour & Information Technology, 23(6), 435-446.

9. Barde, A., Ward, M., Helton, W. S., Billinghurst, M., & Lee, G. (2016, September). Attention Redirection Using Binaurally Spatialised Cues Delivered Over a Bone Conduction Headset. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 60, No. 1, pp. 1534-1538). Sage CA: Los Angeles, CA: SAGE Publications.
10. Spagnol, S., Wersényi, G., Bujacz, M., Bălan, O., Herrera Martínez, M., Moldoveanu, A., & Unnthorsson, R. (2018). Current use and future perspectives of spatial audio technologies in electronic travel aids. Wireless Communications and Mobile Computing, 2018.
11. Mandanici, M., Rodà, A., & Ricca, M. (2018). The Task of Walking Straight as an Interactive Serious Game for Blind Children.
12. Parviainen, J. (2011). Dwelling in the virtual sonic environment: a phenomenological analysis of dancers' learning processes. The European Legacy, 16(5), 633-647.
13. Massen, C., & Prinz, W. (2009). Movements, actions and tool-use actions: an ideomotor approach to imitation. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1528), 2349-2358.
14. Kourtis, D., Sebanz, N., & Knoblich, G. (2010). Favouritism in the motor system: social interaction modulates action simulation. Biology letters, 6(6), 758-761.
15. Brandenburg, K., Brix, S., & Sporer, T. (2009, May). Wave field synthesis. In 2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (pp. 1-4). IEEE.
16. De Vries, D., & Boone, M. M. (1999). Wave field synthesis and analysis using array technology. In Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452) (pp. 15-18). IEEE

Embodiment and Interaction as Common Ground for Emotional Experience in Music

Hiroko Terasawa[1], Reiko Hoshi-Shiba[2], and Kiyoshi Furukawa[3]

[1] University of Tsukuba
[2] The University of Tokyo
[3] Tokyo University of the Arts
terasawa@slis.tsukuba.ac.jp

Abstract. Physical activities and social interaction play crucial roles in the experience of musical emotion. This paper first discusses the physiological functions and social aspects related to music. Upon the observations we introduce the network structure underlying the emotional experience in traditional music, and extend the framework to contemporary computer music, in which computer-aided embodiment, interaction, and collaboration provide emotional experiences to the players.

Keywords: Musical emotion, interaction, embodiment, participation.

1 Introduction

Physical movements and communication are inseparable from music, coexisting in our musical experience in everyday life. These are both seen when a baby is cuddled and sung a lullaby, when children sing while playing together, when adults perform music in an ensemble, and when a community of people share music during a festive occasion. The most vivid musical emotions that we experience exist within the context of physical movement and communication.

Computer music works in embodied, interactive, and collaborative forms are often experienced emotionally. Compared to works composed as purely auditory expressions, embodied and interactive computer music seems to add another dimension of fascination to players and audiences. This paper discusses such social and physiological elements in music as a foundation for musical emotion.

Classic studies in musical emotion often employ the viewpoint that musical emotion is considered personal responses to "music as auditory stimuli" or "organization or structure of sounds in music." Psychological experiments on music emotion strictly control the listener's bodily actions and the external listening environment to ensure the reproducibility and reliability required for sound scientific research. These studies provide much information on the perception and cognition of music on an individual basis, which is fundamental to understanding musical emotion. However, they do not directly explain the dynamic and lively musical emotion that we experience daily. According to Johnson [1], "*the experience of sitting quietly in a chair and listening to music is almost unnatural, for our bodies want to move with the music*." Another often neglected issue is the social aspect in music; DeNora vividly illustrates how human relationships influence everyday musical experience [2]. This paper aims to provide some theoretical thoughts on the real and non-measurable experience of musical emotion in our actual life.

This paper attempts to describe that (1) musical emotion becomes more precise and enriched with physical and social elements in music, (2) physical and social aspects of

musical emotion can be described in the form of network structure, and (3) emotional experience in embodied, interactive, and collaborative computer music pertains to the same framework.

Before starting our discussion of shared musical emotion, we need to define music, emotion, and musical emotion. Answers to the question "what is music?" are dependent on the cultural and social background of individuals, and many individuals have very strong opinions and images about music [3]. In this discourse, we focus on the embodied and interactive aspects of musical experiences, which is shared by both musicians and non-musicians; therefore, it is appropriate to refine the definitions from standard language dictionaries, assuming that they reflect the ideas and images of music of ordinary people, not just artists or philosophers:

Music is artistic creation, sonic production (not necessarily considered to be artistic), and cultural and social behavior that combines the sounds produced by humans using the elements and styles including (but not limited to) rhythm, melody, harmony, chanting, etc., which are distinct from speech.

Emotion can have diverse definitions; however, this paper followed the definition from the "Handbook of Music and Emotion" by Sloboda and Juslin [4]:

Emotions are relatively brief, intense, and rapidly changing responses to potentially important events (subjective challenges or opportunities) in the external or internal environment. They are usually of a social nature, which involves a number of subcomponents (cognitive changes, subjective feelings, expressive behavior, and action tendencies) that are more or less "synchronized."

Musical emotion is then defined as *"emotions that were somehow induced by music"* [4]. Along with the above definition of emotion, when we experience musical emotion, "the potentially important events (subjective challenges or opportunities)" are music. When considering musical emotion, music should be regarded not only as stimuli from the "external environment" but also as agency in the internal environment (physiological and psychological conditions) and as a social medium.

2 Emotional information in musical activities

2.1 Embodiment: The Physiological and Physical Ground of Musical Emotion

The physiological and physical elements related to musical emotion vary from unconscious and unintentional physiological functions to conscious and intentional actions of musical performance. According to the James-Lange theory of emotion (i.e., the physiological condition of the human body determines the experience of emotion) [5], we can presume that the physiological and physical aspects of music determine the experience of musical emotion.

Studies on physiological reactions related to music include dopamine release matching the highest arousal of musical emotion [6], presence of mirror neurons [7], mirror system synchronized with musical performance gestures [8], rhythmic synchronization [9], rhythmic prediction processes in brain waves [10], and chills [11]. Physiological studies related with music are currently conducted in many institutions, and more outcomes are expected to appear. These basic motor functions and physiological responses to music are unconsciously produced and hard to control. Therefore, it is difficult to estimate how much they contribute to the physical context of our musical experience. However, since they are unavoidable and unconscious, it is reasonable to assume that they always exist and form the foundation of conscious, active motions in musical experiences.

Performance skill relies on motor skill acquisition, which affects the brain structure while training for a long time. Skilled pianists spend a long time developing and maintaining their performance skill [12, 13]. Skilled performers who started their training at an early age have a significantly larger area corresponding to their hands in the motor cortex in their brain than non-musicians do [14]; they have an enlarged genu, the signal pathway area in brain related to motor functions [15], and a stronger connection between the neural pathway for music and listening and the pathway for hand motions [16]. Overall, performance training during early childhood enlarges the functional area of the brain's motor skills, stimulates the activation of the area, and creates a neural pathway for performance. An artistic and expressive performance requires precise control of physical movements, and acquiring such controls demands rigorous training and re-wiring of neural pathways of the brain simultaneously. We could say that using our body intensively for making music transforms our brain and presumably increases the sensitivity, accuracy, and resolution in both music listening and making.

2.2 Collaboration, Participation, and Interaction: The Social Element of Musical Emotion

Social aspects of musical emotion are often anecdotal, yet I trust that many people feel that live musical experiences in groups tend to be more emotional than experiencing music in isolation.

Most Western traditional music, which is expected to have a composer, players, and listeners, is a social behavior delivering sounds from person to person, and music performances require coordination and communication among performers in a multimodal way [17]. Luck and Toiviainen studied the synchronization between conducting gestures and performing music using a motion-capture system, finding that the performers synchronized most precisely at parts with a gradually reducing or very fast tempo [18]. Other studies report that visual observations of conducting [19] or performing gestures [20] can deliver emotional expressions to participants regardless of their musical experience. These studies show that musical emotion is communicated not only in the auditory domain but also in multimodality including motions and visions.

Many cultures in the world share dynamic and strong emotions through music. In many cases, such as rituals and festivities, players and listeners are not differentiated. People play, sing, dance, and share in very strong emotions together, sometimes reaching high degrees of excitement. In these situations, active performance, listening, and physical motion occur at the same time, and music is felt as a shared dynamic experience, providing strong emotional bonding.

This kind of musical experience exists not only in traditional rituals but also in modern cultures. According to Csikszentmihalyi [21]:

"The audiences at today's live performances, such as rock concerts, continue to partake in some degree in these ritual elements; there are few other occasions when large numbers of people witness the same event together, think and feel the same things, and process the same information. Such joint participation produces (…) the sense that one belongs to a group with a concrete, real existence."

Performance, especially improvisation, requires strong attention and focus and tends to promote dynamic communication of musical emotion [22]. This paradigm of group improvisation in motion is employed in many multimedia sound-art installations.

Delivery and sharing of musical emotion rely not only on sonic information but also on sharing motions, gestures, and visions; musical and performance expressions can deliver emotions using sound and music. Yet, from the viewpoint of the James-Lange theory of emotion, if people move their bodies together to play music, they

share the physical condition, resulting in sharing the emotional condition. That means that sharing motion along with music accelerates emotion sharing among people. The emotion experienced may not be exactly same, yet "the dynamic contour of musical emotion" is shared by people in a synchronized manner.

3 A Network Model of Musical Emotion

This section describes a model for the communication of musical emotion. In this model, the communication of musical emotion is described as a network of human agents with an internal system and an interface of perception and expression. Maturana and Varela describe a schematic of connected agents with internal state which are affected by a surrounding environment [23]. A similar model of perception and motor function sharing is also presented in Benzon's literature [22]. These literatures describe the physiological and psychological conditions between coupled agents (i.e., interacting people), but do not discuss the emotion-sharing mechanism in musical situations in detail. Given this context, we wish to extend the discussion on emotional communication in music in this section. This study regards musical experience as a communication of musical emotion, and music as a medium of emotion. We aim to explain that musical emotion sharing is founded upon the physical and social aspects.

3.1 Physiological Model for Emotional System

The human body is the fundamental agent in the communication of musical emotion. Humans relate to the external world by perception (introverted) and expression (extroverted). The external, incoming information is perceived and modifies the already existing psychological and physiological conditions. Let us name this integrated entity of psychological and physiological conditions as one's "internal state." Inside the body, neural responses including brain activities and hormonal secretions occur, changing the internal state dynamically. Extroverted phenomena such as physiological responses, motions and gestures, facial expressions, performance, and voice transmit the internal state to the outside world. We can assume that emotion exists in a manner that integrates the conscious and unconscious elements in the brain. The internal system should have a function to integrate the conscious and unconscious elements as a preparatory stage for emotion. A simplified model for this dynamic system is drawn in Figure 1.

.2 Bi-directional Communication of Musical Emotion

We first consider the simplest case in which two humans share emotion. This mutual sharing of musical emotion is described in Figure 2. The emotion that exists inside the internal system of performer A is expressed to the outside world with the expressive interface, and the expressed emotional information is delivered to performer B's internal system via their perceptual interface.

3

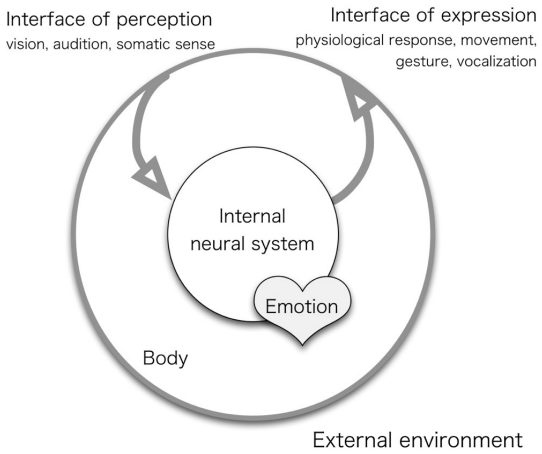


Fig. 1. The model of the emotional system. Humans are viewed as a physical system with the interface of perception/expression and the internal system that encloses emotion. The internal system consists of the central nervous system including the mirror system; the peripheral nervous system including autonomic, somatic, and motor nervous systems; and the endocrine system.

In this model, music is a medium for emotional expression and perception. However, the medium of emotional delivery is not limited to music as a sound, but other accompanying factors of music such as gestures, movements, and visual presentations. The musical information (i.e., solid lines in Fig. 2), extra-musical information (i.e., dashed lines in Fig. 2), and internal state of the perceiving person affect the emotion in a musical experience as a whole.

Bi-directional communication of musical emotion is mutual: performer B's reaction and expression also affect performer A's emotion dynamically. In this way, the feedback loop for the emotion sharing is established.

With this loop structure, emotion can dynamically travel between two people. When this transmission of emotion happens continuously and it is synchronized with music, the sharing of musical emotion described in Section 2 occurs. When musical emotion is shared, each participant understands and responds to each other's emotion, and they progressively exchange their emotion through music. This process enables the communal production and synchronized experience of the dynamic contour of musical emotion.

This structure is true whether the two people are both performers or one is a performer and the other a listener. According to Small, *"musicking"* refers to participating in a musical performance, including both of playing and listening [24]. Small emphasizes rather passive attitudes of listeners in classical music concerts in the main chapters of his book: in such situations, the arrow from the listener to the performer in this model is very thin. However, in other genres, active participation of listeners is welcomed. In such settings, performers often change their expression based on the listeners' reactions, and listeners take an active part in a musical performance, thickening the arrow from the lister to the performer in the model.

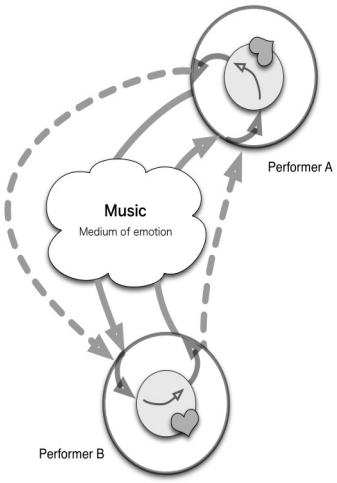


Fig. 2. Communication of musical emotion is established by the feedback at the internal system described by the arrow in the internal system. Musical emotion travels between two people via the medium of musical information (solid line) and extra-musical information (dotted line, e.g., bodily information).

3.3 Multi-directional Communication of Musical Emotion

The experience of dynamic musical emotion shared by a group of people is described in Figure 3. With "song and dance" experiences and group improvisations, participants formulate a network and intervene in musical production. In such musical performances, each person connects to everyone else's emotion using the interface and internal system and actively participates in the music, creating the musical emotion together. This is an N:N dynamic communication system that uses musical and extra-musical information as a medium of emotion.

Some other cases may be regarded as 1:N communication: for example, a live rock or concert performance by a charismatic musician where audiences share overwhelming emotions. Musicians use music and physical actions as their medium, efficiently approaching the audience's interface and internal system and producing very strong emotional responses among many people. However, genuine performances are not unilaterally offered; great musicians promptly sense audiences' vocal and gestural reactions, and deliver dynamic and delicate adjustments in their musical expressions.

In this model, musical and extra-musical information are emphasized with a solid line and a dashed line. Each of the connections represented with these lines can be stronger or weaker, allowing the communication to be distributed unequally, or some functions to be used more intensively.

Overall, our musical experience is a multimodal process that integrates the auditory, visual, and somatosensory interfaces and exploits the full function of the internal system. The produced music is a fruit of this dynamically collaborative and multimodal process and reflects the dynamic trajectory of musical emotion shared by the participants.

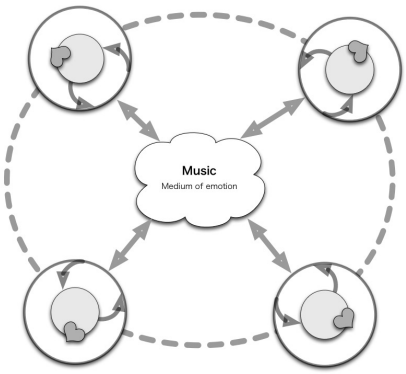


Fig. 3. Dynamic communication of musical emotion of a group of people. While the participants interactively intervene into music, they share the musical emotion with music (solid line) and extra-musical information (dotted line) as the medium. The dynamical transformation of music induces further transformation of musical emotion.

4 Embodiment and Interaction in Computer Music

Embodied and interactive works in computer music often offer emotional experiences to both audiences and performers. In this section, we attempt to relate these emerging musical styles to the network model of musical emotion.

4.1 NIMEs and Computer Music

New interfaces for musical expressions (NIMEs) offer physical controls for computer music performances in real time. The means of acquiring control signals vary from sensors to physiological signals (such as EMG and EEG).

One of the earliest NIMEs was a radio baton invented by Max Mathews [25], an interface that allows control of the rhythm and dynamics of music. Other early sensor-based NIMEs investigated the direction of computer music being "performed" with the human body [26, 27] and gave a strong impression to the audiences that NIMEs are equally playable and playful as traditional musical instruments.

This direction of embodiment became even more progressive by transforming the human body into instrument. Some musicians used physiological signals such as EMG (electromyography) and motion-related signals (e.g., using accelerometer) to control music dynamically [28, 29]. Their ways to control music were more sonification-oriented, rather than being MIDI-based. The message of the "human body as an instrument" attracted audiences. Musical performances that use EEG (electroencephalography) also provide similarly emotional, often intimate experiences to audiences [30, 31].

Although we tend to focus on the technical innovations when considering NIMEs, their emotional offerings cannot be ignored. Both performers, who are often inventors themselves, and audiences are focused and engaged in the musical performance with a full range of emotions.

The emotional quality experienced in NIME performances is equal to that of a classical music performance, and the same structure of musical-emotion communication can be observed. The performers communicate with audiences using

musical and extra-musical information. Both the real-time music rendering and visual presentation of the performing gestures enable the immediate delivery of musical emotion from the performer to the audiences. The audiences provide reactions to the music, and the performers can reflect that into the music. This mutual communication of musical emotion is present in NIME performances.

4.2 Laptop Orchestra

Laptop orchestras are practiced in many universities, and their courses are popular among students [32-34]. The performers of laptop orchestras have to observe, react, and participate in musical performances with other performers. Many repertoires require the performers to not only play but also improvise according to the music. This framework also allows an intensive communication of musical emotions among performers themselves and with audiences. The musical emotion is expressed in terms of musical and extra-musical information (in this case, gestures, motions, shared-codes, messages over networks, etc.) and performers have to musically respond in time with gestures, which are captured by sensors [35]; these musical action and responses form the network structure of emotional communication. This performative communication provides a strong sense of unity to the performers and audiences, making the laptop orchestra an emotional experience to both performers and audiences [36].

4.3 Networked Performances

The idea that people perform music together from a distance is highly attractive and has fascinated many musicians. Some network-connected performances are conducted in concert form with audiences [37, 38] while others are conducted as a gig among musicians without expecting particular audiences [39]. These performances involve singing, playing traditional instruments, and performing NIMEs or mobile apps. Again, the network structure of musical emotion communication is present and musicians are having a deeply emotional experience playing music. The main difference from traditional music is that the medium and pathway for musical and extra-musical information are now on Internet. However, the emotional experience in the networked music is fundamentally similar to that in the traditional music making, or perhaps augmented with the excitement to perform music together with unexpectedly distant musicians.

4.4 Collaborative and Participatory Paradigms

Collaboration and participation have become prominent factors in computer music. Many platforms and interfaces that enable collaborative creation of music have been proposed [40-43]. Installation, sound-art, and audience-participation performances are also very popular [44-47]. These systems function as a foundation for communicating musical emotion; the systems can reflect the intentions and emotions of participants in real-time by means of sound, vision, gestures, motions, and physiological conditions. Jo et al. describe the participatory design of music based on community, where people intensively focus on the process of creating new music [48]. With these participatory frameworks, participants can exchange musical ideas and emotions forming the network structure, and their musical performances offer emotional experiences to both participants and audiences.

5 Discussion

5.1 Audio as a Fluid Medium

In this model of communication of emotional sharing, music is regarded as the medium of emotional expression. The acoustic quality of music promises immediate intersubjectivity by mixing sounds.

When multiple people express their emotion, the expressions do not easily merge if they are not in the acoustic medium. Expressions by a solid object or physical motion exist individually and do not naturally blend together. Expressions intermix when the information is represented in a "mixable" medium such as sound, light, and liquid. By choosing sound as the expressive medium, music allows people to fuse their expressive information and construct an auditory scene where people can detect auditory objects if the sound production is well controlled [49].

The mixing of the sound medium is characteristic as multiple sounds intermix in real time with very little directive influence, creating an acoustic field. Each participant in the acoustic field, even if the person produces a different sound than the other people, is exposed to the same acoustic event at the same time as the others, which is constructed of many different sounds contributed to by everyone in the field.

This mixing/fusing quality of music offers the physical basis for the immediate emotion sharing in the communication of musical emotion. Therefore, music is a physically intersubjective medium. Such a quality enables the intersubjective experience of musical emotion among multiple participants.

5.2 Discovering New Values in Music

The perspective of dynamic and collaborative communication of musical emotion enables us to consider new musical values. Interactive sound art, embodied musical instruments and NIMEs, participatory music, and collaborative computer music performances, as illustrated in Section 4, are emerging genres that progressively expanded after the 2000s with the development of computer and internet technologies. These new musical works are often presented in computer music and media art events such as ICMC, NIME, Ars Electronica, and SIGGRAPH. These repertoires do not necessarily pursue the traditional values in common Western music such as abstract musical expressions [50], but rather exploit the dynamic and embodied experiences aided by technologies and the musical emotions themselves as media of expression.

The values of traditional Western music are rooted in the traditional Western music theory. In order to focus on the value of music as an abstract art, abstract descriptions (e.g., score-based notation) are essential. Moreover, the history of music theory and music analysis based on scores formulated the values of traditional Western music.

We believe the above-mentioned new musical works should be evaluated in terms of the theory of musical emotion rather than the values of traditional Western music. Our model and theory for musical emotion enables the evaluation of contemporary works in terms of contemporary perspectives. Such a discourse will develop a new system of musical works and theory and will eventually help us discover new musical values.

Interactions and embodiment in music have not been fully considered in music theory; consequently, some musical aspects such as performance and participation have plenty of room for further theoretical development. Theorization of this research area and enfolding it into the system of musical theory is beneficial for the further development of music theory and musicology. Our discourse on musical emotion in

physical and social contexts, along with the related studies to unveil the gestural, emotional, and social circumstances of new music [51-55], hopefully contributes to the basis for a new theory for music.

6 Conclusion

In this paper, we described the common structure found in emotional communications in both traditional and new music using a network model of musical emotion. The exchange of emotion via musical (i.e., sounds) and extra-musical information (i.e., gestures, visuals, motions, messages, etc.) is the key for live music experiences and is one of the sources of focus, fascination, and excitement in live performances of new music. Embodiment and interactions are essential factors for this structure. We think this structure itself is a musical heritage that is common for the music of various styles and cultures, including computer music.

References

1. Johnson, M.: The meaning of the body: Aesthetics of human understanding. University of Chicago Press (2007)
2. DeNora, T: Music in everyday life. Cambridge University Press (2000)
3. Nettle, B.: "Music." In Grove Music Online, Oxford Music Online. http://www.oxfordmusiconline.com/subscriber/article/grove/music/40476
4. Sloboda, J.A., Juslin, P.N.: At the interface between the inner and outer world: Psychological perspectives. In: Juslin, P.N., Sloboda, J.A. (eds.), Handbook of Music and Emotion, Theory, Research, Application pp. 73-98. Oxford: Oxford University Press (2010)
5. Cannon, W.: The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. Am J Psychol. 39, 106–124 (1927)
6. Salimpoor, V.N., Benovoy, M., Larcher, K., Dagher, A., Zatorre, R.J.: Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nat Neurosci. 14(2), 257–262 (2011)
7. Rizolatti, G., Craighero, L.: The mirror-neuron system. Annu Rev Neurosci. 27, 169-192 (2004)
8. Zattore, R., Chen, J., Penhune, V.: When the brain plays music: Auditory-motor interactions in music perception. Nat Rev Neurosci. 6, 692–695 (2007)
9. Merker, B.H., Madison, G.S., Eckerdal, P.: On the role and origin of isochrony in human rhythmic entrainment. Cortex. 45, 4–17 (2009)
10. Iversen, J.R., Repp, B.H., Patel, A.D. Top-down control of rhythm perception modulates early auditory responses. Ann N Y Acad Sci. 1169(1), 58–73 (2009)
11. Huron, D.: Sweet anticipation. MIT press (2007)
12. Williamon, A, Valentine, E.: Quantity and quality of musical practice as predictors of performance quality. Br J Psychol. 91, 353–376 (2000)
13. Jabusch, H.C., Alpers, H., Kopiez, R., Altenmüller, E.: The influence of practice on the development of motor skills in pianists: A longitudinal study in a selected motor task. Hum Mov Sci. 28, 74–84 (2009)
14. Amunts, K., Schlaug, G., Jäncke, L., Steinmetz, H., Schleicher, A., Dabringhaus, A., Zilles, K.: Motor cortex and hand motor skills: Structural compliance in the human brain. Hum Brain Mapp. 5, 206–215 (1997)
15. Schlaug, G., Jäncke, L., Huang, Y., Staiger, J.F., Steinmetz, H.: Increased corpus callosum size in musicians. Neuropsychologia. 33(8), 1047–1055 (1995)
16. Bangert, M., Peschel, T., Schlaug, G., Rotte, M., Drescher, D., Hinrichs, H., Heinze, H.J., Altenmüller, E.: Shared networks for auditory and motor processing in professional pianists: Evidence from fMRI conjunction. NeuroImage. 30, 917–926 (2006)
17. Turino, T.: Music as Social Life: The Politics of Participation. University of Chicago Press (2008)

18. Luck, G., Toiviainen, P.: Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. Music Percept: An Interdisciplinary Journal. 24(2), 189–200 (2006)
19. Luck, G., Toiviainen, P., Thompson, M.R.: Perception of expression in conductors' gestures: A continuous response study. Music Percept: An Interdisciplinary Journal. 28(1), 47–57 (2010)
20. Davidson, J.W.: Visual perception of performance manner in the movements of solo musicians. Psychol Music. 21(2), 103–113 (1993)
21. Csikszentmihalyi, M.: Flow: The psychology of optimal experience. Harper and Row (1990)
22. Benson, W.L.: Beethoven's Anvil. Basic Books (2001)
23. Maturana, H., Varela, F.: The tree of knowledge. Shambhala (1992)
24. Small. C.: Musicking. Wesleyan University Press (1998)
25. Mathews, M. V.: The Radio Baton and Conductor Program, Or: Pitch, the Most Important and Least Expressive Part of Music. Comp Music J. 15(4), 37–46 (1991)
26. Gurevich, M., von Muehlen, S.: The Accordiatron: A MIDI controller for interactive music. In: Proceedings of the international conference on New Interfaces for Musical Expression pp. 1–3 (2001)
27. Wilkerson, C., Ng, C., Serafin, S.: The Mutha Rubboard Contoller: Interactive Heritage. In: Proceedings of the international conference on New Interfaces for Musical Expression. pp. 82–85 (2002)
28. Tanaka, A.: Musical Performance Practice on Sensor-based Instruments Trends. In: Gestural Control of Music, Wanderley M.M., Battier, M. (eds.) e-book: http://www.idmil.org/projects/trends
29. Nagashima, Y: Bio-sensing systems and bio-feedback systems for interactive media arts. In: Proceedings of the international conference on New Interfaces for Musical Expression, pp. 48–53 (2003)
30. Leslie, G.: Vessels: a brain-body performance. Musical composition (2015)
31. Hamano, T., Rutkowski, T., Terasawa, H., Okanoya, K., Furukawa, K.: Generating an integrated musical expression with a brain–computer interface. In: Proceedings of the international conference on New Interfaces for Musical Expression, pp. 49–54 (2001)
32. Trueman, D., Cook, P.R., Smallwood, S., Wang, G.: PLOrk: Princeton Laptop Orchestra, Year 1. In: Proceedings of the International Computer Music Conference. New Orleans (2006)
33. Trueman, D.: Why a laptop orchestra? Organised Sound 12(2), 171–179 Cambridge University Press (2007)
34. Wang, G., Bryan, N., Oh, J., Hamilton, R.: Stanford Laptop Orchestra (SLOrk). In: Proceedings of the International Computer Music Conference. Montreal (2009)
35. Wang, G.: Artful design. Stanford University Press (2018)
36. Pardes A.: The Aural Magic of Stanford's Laptop Orchestra. Wired Magazine: https://www.wired.com/story/stanford-laptop-orchestra-tenth-anniversary-concert/ (2018. URL retrieved in July 2019)
37. Cáceres, JP., Hamilton, R., Iyer, D., Chafe, C., Wang, G.: To the edge with China: Explorations in network performance. In: ARTECH: Proceedings of the 4th International Conference on Digital Arts, pp. 61–66 (2008)
38. Hamilton R., Caceres, JP., Nanou, C., Platz, C.: Multi-modal musical environments for mixed-reality performance. Journal on Multimodal User Interfaces. 4(3-4), pp. 147–156 (2011)
39. Wang, G., Essl, G., Smith, J., Salazar, S., Cook, P. R., Hamilton, R., Fiebrink, R., Berger, J., Zhu, D., Ljungstrom, M., Berry, A., Wu, J., Kirk, T., Berger, E., Segal, J.: Smule= Sonic Media: An Intersection of the Mobile, Musical, and Social. In: Proceedings of International Computer Music Conference, pp. 283–286 (2009)
40. Leslie, G., Mullen, T.: MoodMixer: EEG-based Collaborative Sonification. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 296–299 (2011)
41. Barraclough, T., Murphy, J., Kapur, A.: New Open-Source Interfaces for Group-Based Participatory Performance of Live Electronic Music. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 155 – 158 (2014)
42. Lee, S. W., Essl G.: Communication, Control, and State Sharing in Networked Collaborative Live Coding. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 263–268 (2014)

43. Hazar Emre Tez, Nick Bryan-Kinns: Exploring the Effect of Interface Constraints on Live Collaborative Music Improvisation. NIME, pp. 342–347 (2017)
44. Sergi Jordà: Sonigraphical Instruments: From FMOL to the reacTable.* NIME 3, pp. 70–76
45. Matsumura, S., Arakawa, C.: Hop step junk: sonic visualization using footsteps. In: Proceedings of the international conference on new interfaces for musical expression, Vancouver, BC, Canada, pp. 273–273 (2005)
46. Taylor, R., Schofield, G., Shearer, J., Boulanger, P., Wallace, J., Olivier, P.: Humanaquarium: A Participatory Performance System. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 440–443 (2010)
47. Lind, A., Nylén, D.: Mapping Everyday Objects to Digital Materiality in The Wheel Quintet: Polytempic Music and Participatory Art. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 84–89 (2016)
48. Jo, K., Parkinson, A., Tanaka, A.: Workshopping Participation in Music. Organised Sound 18(3), pp. 282-291 (2013)
49. Bregman, A.S.: Auditory scene analysis: The perceptual organization of sound. MIT press (1994)
50. Kania, A.: The philosophy of music. Stanford Encyclopedia of Philosophy https://plato.stanford.edu/archives/fall2017/entries/music/ (2017)
51. Ferguson, S., Schubert, E., Stevens, C.: Movement in a contemporary dance work and its relation to continuous emotional response. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 481–484 (2010)
52. Jaimovich, J., Ortiz, M., Coghlan, N. R., Knapp, B.: The Emotion in Motion Experiment: Using an Interactive Installation as a Means for Understanding Emotional Response to Music. In: Proceedings of international conference on New Interfaces for Musical Expression (2012)
53. Graham Booth, Michael Gurevich: Collaborative composition and socially constructed instruments: Ensemble laptop performance through the lens of ethnography. NIME (2012)
54. Morgan, E., Gunes, H., Bryan-Kinns, N.: Instrumenting the Interaction: Affective and Psychophysiological Features of Live Collaborative Musical Improvisation. In: Proceedings of international conference on New Interfaces for Musical Expression, pp. 23–28 (2014)
55. Leslie, G. Ojeda, A. Makeig, S.: Measuring musical engagement using expressive movement and EEG brain dynamics. Psychomusicology: Music, Mind, and Brain 24(1), pp. 75-91 (2014)

Computer Generation and Perception Evaluation of Music-Emotion Associations

Mariana Seiça, Ana Rodrigues, F. Amílcar Cardoso, Pedro Martins, and Penousal Machado

CISUC, Department of Informatics Engineering, University of Coimbra
{marianac, anatr, pjmm, amilcar, machado}@dei.uc.pt

Abstract. Music is intertwined with human emotions as an artistic form with expressive qualities. We present a pilot study of music-emotion associations based on a generative system, which produces parameter-based music to represent four emotions: Happiness, Sadness, Calm, and Anger. To study the perceptual relevance of each parameter, we performed a series of user tests where participants explored multiple combinations of musical parameters to reach a representation for each emotion. Results were compared with the ones from previous studies and empirical experiments proposed by other authors, which gave us a starting point to evaluate each association and discover new possible connections. Although most of the associations were confirmed, a few discrepancies were found, such as the user preference for low pitch in Anger over the expected high pitch. These findings provide better insight and validation of the relationship between music and emotions, and thus a starting point to explore novel representations.

Keywords: algorithmic composition, generative music, music-emotion associations, emotion perception

1 Introduction

Representing emotions through a computational artifact dates from Picard's concept of *Affective Computing*, i.e., how can computers express/recognize affect and gain the "ability to "have emotions" [1]. Picard defended that emotions are an important part of human cognition and perception, and thus have a prominent role in assisting people with computational systems. Music is by nature a subjective field, which resonates with the subjectivity of emotions, and may be the reason why emotions have been so largely used to manipulate music computationally. The most used emotion model for music experiments is the dimensional model proposed by Russell, where emotions are distributed in a two-dimensional space and split by the dimensions of arousal and valence [2, 3]. Based on this model, we chose a set of four discrete emotions evenly distributed in the 2D space (see Fig.1) to ensure a validation of the most perceptually-relevant parameters.

Despite the potential of using emotions to algorithmically compose music, research on this subject is usually limited to a small number of parameters such

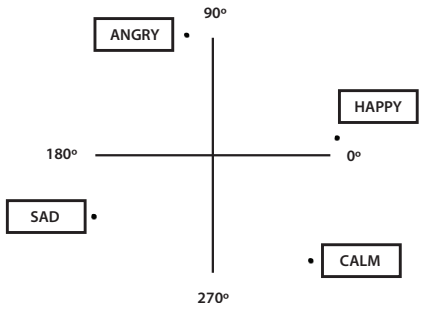


Fig. 1. Choice of emotions based on Russell's Circumplex Model of Affect [2]

as pitch and loudness [4]. In this study, we try to overcome this limitation by developing a computational artifact based on seven musical parameters: harmony, tempo, pitch, melody direction, articulation, melody rhythm, and loudness. To confirm and validate the relationships between these music-emotion associations, we performed a pilot user study, comparing the preferences of each participant to the literature findings on the subject. Our contribution lies in providing a better understanding and support of findings on emotion perception of musical parameters and their respective impact in representing a set of emotions, through the improvement and expansion of a previous generative system [5].

The remainder of this paper is organized as follows. Section 2 comprises an overview of music generation systems, and studies about the association between music and emotions. Section 3 details the improvements made to the previous developed system. Section 4 reports the conducted evaluation, whereas Section 5 provides an analysis and discussion of the results as well as user feedback. Finally, Section 6 draws general conclusions and delineates future work.

2 Related Work

"Music can be used to express emotions more finely than any other language"[1]. But how may these musical characteristics influence the musical forms of expression, and how can computational systems learn this information to produce music based on emotions?

2.1 Music Generation Systems and Music-Emotion Experiments

Algorithmic composition is described as the use of a "formal process to make music with minimal human intervention" [6]. Methodologies used to create automated music range from *stochastic* (Markov chains) and *rule-based*, to *AI* models. A recent work on evolutionary music is Scirea et al.'s *MetaCompose* [7]. With the goal of creating music that can express different mood-states in a dynamic environment, the system generates compositions comprised by "a chord sequence, a melody and an accompaniment"[7], dealing with a set of detailed

musical features, such as harmony, melody, pitch, scale, intensity, timbre, and rhythm.

In the line of rule-based models, Livingstone et al.'s computational music-emotion engine [8] presents a set of rules for the score structure and the performative expression over the arousal-valence model, varying musical parameters like the tempo, mode, loudness, articulation, pitch and others to produce, and specifically induce, certain emotional effects in the listener. The term *Generative Music* was popularized by the composer Brian Eno, creating systems that produce ever-changing ambient music through probabilistic rules, such as his first *Generative Music 1* album using SSEYO Koan Software [9], and his last album *Reflection*, available as an infinite piece through an iOS app [10].

David Cope's *Experiments in Musical Intelligence* system is a reference in computer-aided composition, exploring the concept of *recombinance* through the deconstruction of works of classical Western Music, to find common patterns, simulate compositional styles and discover new combinations [11].

2.2 Music and Emotion Associations

Music perception depends on a combination of factors such as an individual's culture, social context, and personality [12]. Although this is a subject commonly discussed and not fully agreed among authors, Juslin has proposed 7 psychological mechanisms [13] through which music may arouse emotions in the listener.

Table 1. Association between a set of music parameters and emotions

	harmony	tempo	melody pitch	melody direction	melody articulation	melody rhythm***	loudness
HAPPINESS	consonant	fast*	high	ascending*	staccato*	dense	loud*
	20%	100%	24%	32%	64%	/	64%
SADNESS	dissonant	slow	low	descending*	legato	sparse	soft*
	20%	100%	36%	16%	56%	/	100%
CALM	consonant **	slow	high*	ascending**	legato**	sparse	soft**
	4%	12%	4%	4%	4%	/	4%
ANGER	dissonant	fast*	high **	ascending	staccato*	dense	loud
	12%	96%	4%	8%	44%	/	100%

* other characteristics for this parameter were found in the literature / ** few experiments found in the literature
*** not found in the literature, based on previous experiments of Seiça et al. [5]

Experiments conducted on music-emotions associations have generally studied the features of loudness, tempo, and consonance/dissonance of harmonies [4, 14]. For example, consonant harmonies have been associated with positive emotions, and loud loudness and fast tempo with emotions of high arousal [14, 15]. A list of the associations found in the literature [4, 14–19] is summarized in Table 1.

Cultural Aspects: Cultural aspects may impact the way we perceive music. For example, in Western music - the one we are working with - melodies played using notes from a major scale tend to be interpreted as happy, while those played with notes from a minor scale usually sound sad [3, 17]. Additionally, the most common chord trajectories of Western tonal music usually begin by establishing a tonal center or base (tonic), and then step away from the stability using more dissonant chords to build tension, to finally return back to the tonic to create relaxation. As a result, the tonic is often the most frequently played note or the note with the longest duration [19].

3 The System: Affective Music Generation

This work is an improvement of a system previously developed by Seiça et al.[5], whose focus was to musically represent emotions retrieved from Twitter. Built as a rule-based system through probabilities, it was guided by two major musical aspects: harmony and melody. The system worked as a communication between three tools: a Processing sketch for analysing the tweets, a Max patcher for the MIDI generation, and an Ableton live set to produce the final sounds.

According to the probabilities defined for each emotion, the melodic line was shaped based on a melodic scale, defined by the harmonic progression to connect to certain emotional contexts. This setting established the set of possible notes for the melody, specifying the type (scale note, chord note or chromatism), the duration of each note (from whole to eighth notes), and the intervals between them, which shaped the melodic motion and leaps. The harmony was defined through predefined harmonic progressions, which combined the affective nature of different chord natures, played in three possible voicings, and their sequence to represent each emotion. The choice of tone quality had a relatively free structure, where timbres and technical features of synthesized sounds were associated with each emotion, resembling the ambient music genre. For further details on the probabilities implemented and the system, we refer the reader to Seiça et al. [5].

3.1 New Parameters and Variations

In this work, we sought to enhance the system by adding new parameters based on the collected studies, and test the relevance of each musical feature in emotion perception. With this purpose, the values for each parameter were simplified.

For the melody, the type of notes was reduced to two sets (Fig. 2 A): the one originally defined for Anger, and the one for Happiness and Sadness. The same was applied to the rhythm (Fig. 2 B): we maintained two sets of different note durations, a denser and a sparser one, to distinguish the emotions with higher energy (Happiness and Anger), and lower (Sadness and Calm). The intervals were reduced to just one set for all emotions (Fig. 2 C), as it was a parameter which we chose not to evaluate, and hence best to have a neutral role. The melody direction (Fig. 2 D) was one of the new parameters, which defines a tendency for creating ascending or descending melodic lines.

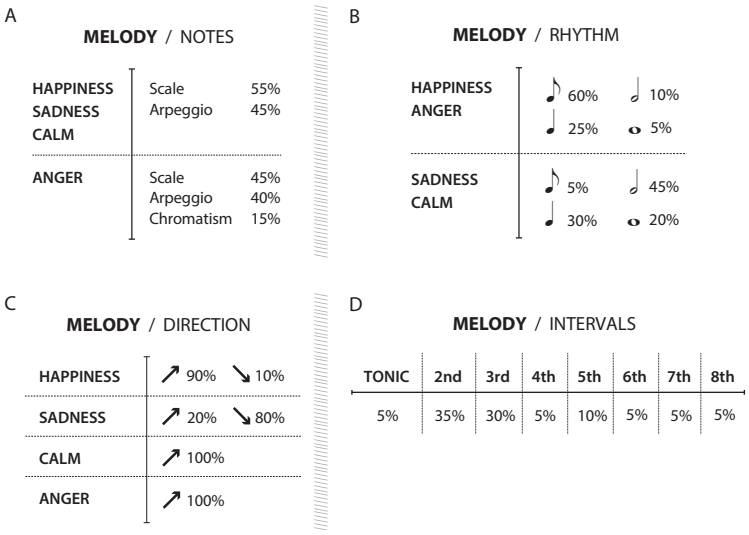


Fig. 2. Probabilities for the type and duration of melody notes, intervals between them and melodic line direction

For the harmony, the progressions were maintained (see examples in Fig. 3 B), and a new parameter was added to control the harmony (Fig. 3 A), playing either a more *consonant* or *dissonant* sound. For the positive emotions (Happiness and Calm), consonance keeps the chords intact, and the dissonance adds notes to the established chords: for example, in a major chord, a minor second, minor third, augmented fourth, augmented fifth or minor seven can be added to the chord structure, all with equal probability. For the negative emotions (Sadness and Anger), dissonance retains the chord structure, and the consonance transforms the dissonant notes of the chords (augmented fourths, augmented fifths and minor fifths) in their consonant counterparts, as perfect fourths and fifths.

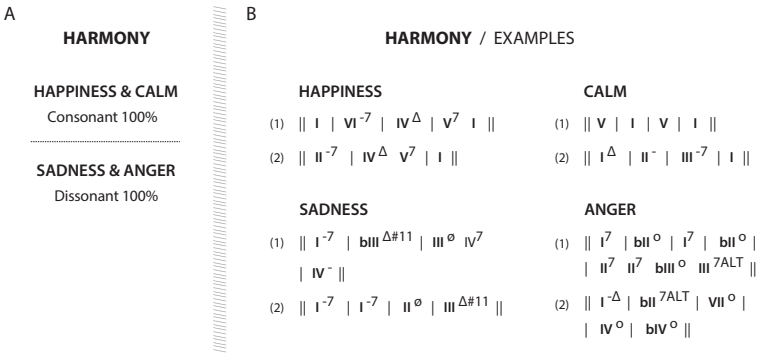


Fig. 3. Examples of harmony progressions and harmony stability for each emotion

The choice for the octave was transformed (Fig. 4 A) to a binary choice of *high* or *low* pitch, which defines the range of possible octaves, with high corresponding to the 4th or 5th, and the low to the 2nd and 3rd. The articulation dynamics was added (Fig. 4 B), which establishes a difference in the harmony and melody dynamics: it can adopt a *legato* style, with each note being played smoothly and connected with each other, or *staccato*, with shorter, detached notes.

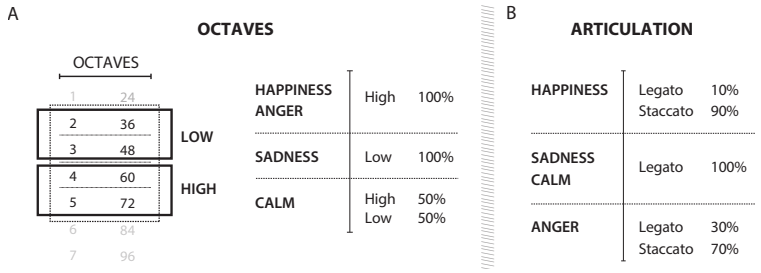


Fig. 4. New octave division, pitch probabilities, and articulation style for each emotion

The control of tempo and loudness, which had already been identified as relevant parameters [5], was implemented as follows. The tempo (Fig. 5 A), measured in BPMs, was defined within three possible range of values: slow (20-75), moderate (76-119) or fast (120-200). Loudness was divided in three MIDI volume levels (Fig. 5 B): soft (44-71), moderate (72-99) and loud (100-127).

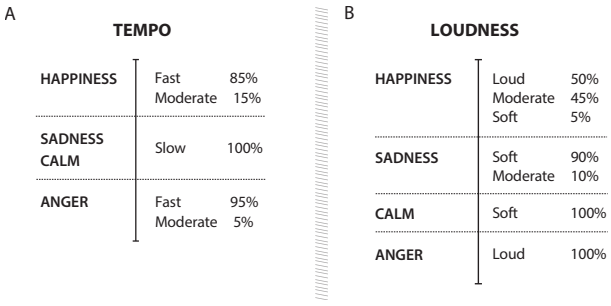


Fig. 5. Probabilities for the tempo and loudness for each emotion

The tone quality was simplified, opting for a piano and a violin, the first to play both the melody and the harmony, and the second for the melody. The choice to reduce the number of timbres to just two instruments, whose sound is familiar and well-recognized, was to balance the tone influence in the emotion association. Therefore, we could focus on evaluating the perceptual weight of the chosen musical parameters with the minimum influence of other characteristics.

4 System Evaluation

We conducted a first set of tests to evaluate the perceptual importance of each musical parameter in representing the four chosen emotions. Seven parameters were tested, according to the number of references in the literature, with the others kept immutable for proper evaluation. The parameters were: harmony, tempo, pitch, melody direction, melody articulation, melody rhythm and loudness. The melody rhythm was an exception we chose to test, despite little mention in the literature review, as it was already implemented in Seiça et al.'s system [5], and we considered it as a relevant feature to evaluate. For each parameter, and based on the literature findings, we established the expected values for each emotion (see Table 1), which would then be compared to the participant's choices.

4.1 Experiment Setup

Twenty participants (12 male and 8 female) took the test. Ages spawned from 22 to 45 years old with an average of 27.8 years and a standard deviation of 4.88. We focused on gathering a balanced set of participants in terms of musical background, distinguishing the ones who have studied music outside the school system and thus have more sensibility to certain musical aspects, from the ones who have not. The tests were performed in person to ensure that the environmental conditions were the same for all the participants.

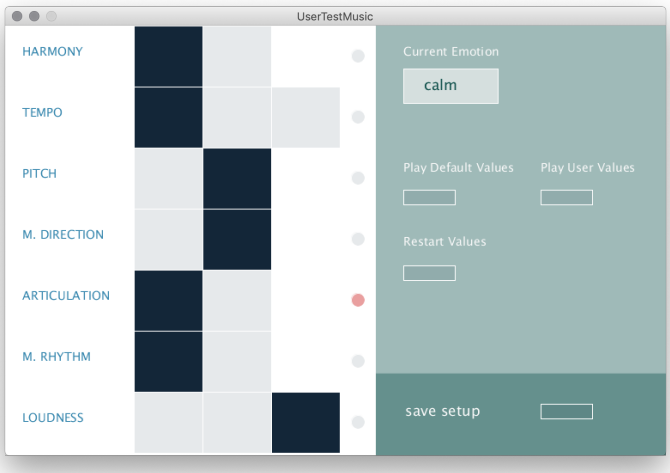


Fig. 6. Interface created for the user tests

We conceived an interface to provide a natural and easy way for participants to interact with our system, which would allow the users to explore the possi-

ble combinations between musical parameters (see Fig. 6) in real time, listen to them, and choose the one that most resembled each emotion. The participants had to aurally interpret the impact of the values for each parameter, with their written designation as the only hint. If the participant, for instance, considered a musical parameter to be irrelevant in the emotion representation - either because he/she could not understand the musical variations or the parameter would not influence the representation (in a positive or negative way) - he had the possibility to mark it through a button for it to be deemed as "indifferent/neutral".

The test would begin with a random combination of musical parameters chosen by the system. This particular choice was to ensure that the participants would not be influenced by the system, and to avoid possible fatigue during the test. The participants also had the possibility, at any point, to listen to the starting point of the system and compare it to their preferences, or even restart the values to the initial set. Once a participant reached the best possible combination, he/she would save the chosen set and evaluate it from 1 to 5 (Likert scale) according to its perception of the emotion representation. In this interval of exchange of feedback there was no music being played, so that the participant could have a moment to refocus and return to a neutral state of mind before the next emotion. This process would repeat through the following order: Happiness, Sadness, Calm and Anger. As the initial set of parameters for each emotion was always random, the order of emotions would not influence its perception: the music could start close or distant from the expected combination of parameters, ensuring a non-biased user's choice.

5 Analysis of Results

For each emotion we have analyzed: (i) the time each participant took to reach a preferred combination; (ii) the satisfaction/resemblance of his/her combination with the represented emotion; and (iii) the relationship between the participant's answers to literature findings. We also analyzed results taking into consideration the musical background of each participant.

Regarding **time**, people with no musical background took longer to grasp the musical parameters and the changes caused by each value. This difference was more pronounced in Happiness, with an average of people with musical background taking 3:37 minutes to reach a desired combination, and the ones without taking 5 minutes. Happiness was followed by Calm, with 1:05 minutes distinguishing the participants with and without musical background, Anger with 30 seconds, and Sadness being the most balanced, with just 3 seconds.

The participants **evaluation/resemblance** of each emotion (see Fig. 7) was measured with a Likert scale. According to this assessment, Sadness was considered the emotion with the best musical representation, with 11 participants having chosen the highest classification (5) of Likert scale. The three remaining emotions had their most popular classification in the fourth value. Above all, people with musical background had a tendency to give a higher classification, which can be justified by a more accurate understanding of music fluctuations.

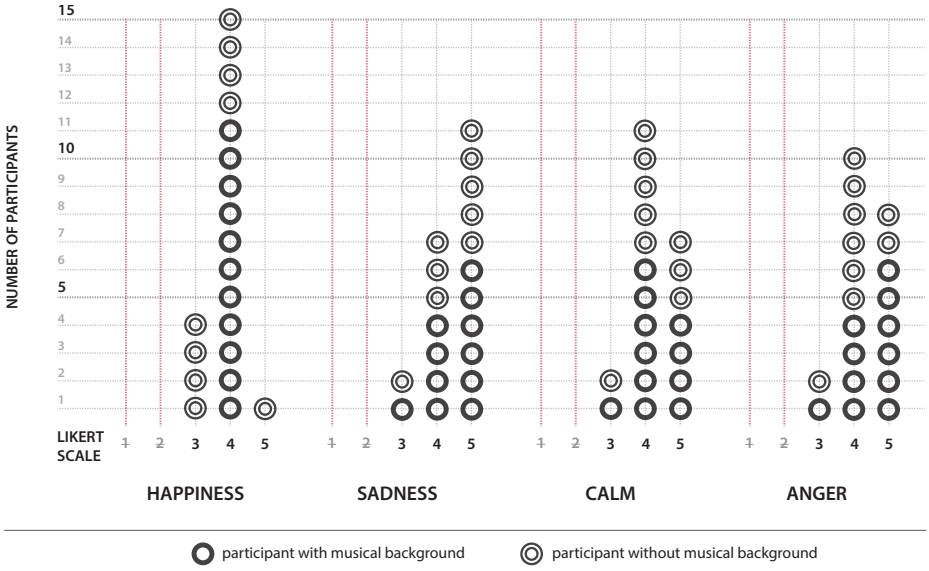


Fig. 7. Likert evaluation for the representation of each emotion

Concerning the **relationship/distance** between participant's preferences and literature findings we performed two types of analysis: (i) distance to the state of art; and (ii) the percentage of correct answers of each music parameter.

The **distance** between the sets of values chosen and the one proposed in the literature is represented in Figure 8. There we can see that despite Happiness having less positive results in the Likert evaluation, had the lowest variation, maintaining the distance between 0 and 3 parameters, which shows a higher correlation with the findings from the literature. Sadness and Anger didn't have a single chosen set equal to the expected, with answers diverging along 6 parameters. Overall, the majority of answers are placed between the distances of 1 and 4, which shows a perceptual tendency and a high correlation with findings from other authors.

As for the **percentage**, we did sum the number of answers that matched the literature findings. Overall, tempo and loudness were the parameters with most answers reflecting these findings, followed by melody articulation, melody rhythm, harmony, melody direction and pitch. One value that stood out was the pitch parameter in Anger, where none of the participants chose the expected value. This may be explained by the scarce number of studies we found in the literature or because it is more perceptually relevant. Happy and Calm were the emotions with more answers matching the literature. Regarding the influence of musical background, there was a fine balance in the answers of participants with and without it, so no major conclusions were drawn in this matter.

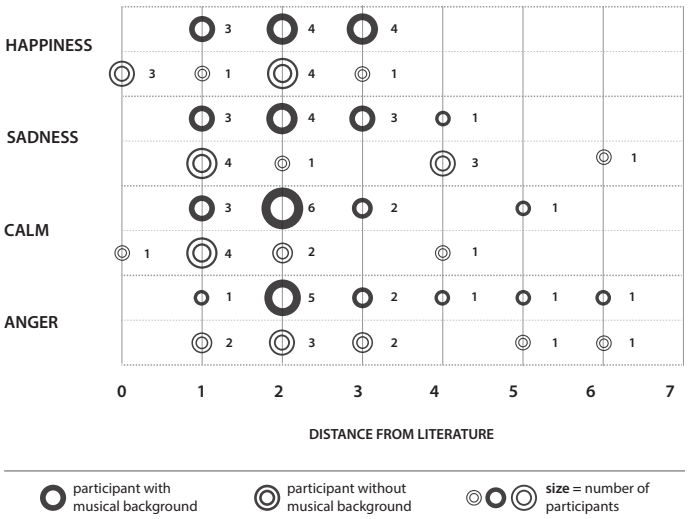


Fig. 8. Distance of the user's choice of parameters to the studies found in the literature. "0" represents the answers matching the literature findings, and "7" no match (all 7 parameters were different)

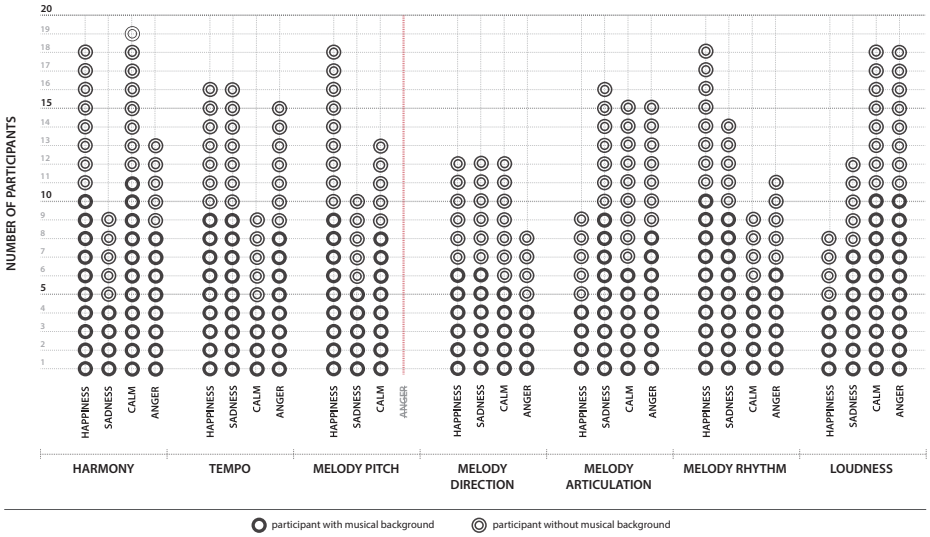


Fig. 9. Distribution of the participants, divided by the musical background, who chose the expected answer concerning each musical parameter

5.1 General Discussion and User Feedback

Overall, the results of the user testing have confirmed a tendency towards the literature. We highlight the difference between participants with and without a

musical background in the time they took to perform the test. Because participants with musical background did grasp musical fluctuations more accurately they were faster concluding the test when compared to users with no musical background. However, as suggested by Juslin [12, 13], there are a series of externals factors that may influence the way we perceive music. For example, we observed that most participants performing our test would compare the system's output to previously known songs or contexts - *evaluative conditioning* [13]. They would then explore different combinations of parameters with the goal of finding the combination that would arouse the same kind of emotions. We also reported that at least four participants made a strong association with cinematic scenarios - *visual imagery* [13]. For instance, they reported to imagine a scenario from a movie while listening to the resulting composition of our system.

Concerning the perceptual relevance of musical parameters, melody direction was considered the feature with less impact in the emotion representation, having been reported by seven participants out of twenty. Melody rhythm and harmony were the succeeding parameters, both with four "indifferent/neutral" answers, followed by tempo with two answers, and articulation with one report.

As for user feedback, five participants reported to feel difficulty in recognizing emotions, as their concepts of emotion "relied a lot on the cultural and musical background". Furthermore, these concepts are volatile, and thus its subjectivity, as there is not just one kind of each emotion: Calm can be happier or sadder, Sadness can be more melancholic, anguishing or even nostalgic.

Regarding musical parameters, four participants reported that Happiness should have a faster rhythm and marked pacing, as they considered it to be a key element in a deeper perception of Happiness. Five participants also noted the lack of intermediate values for some parameters, which would allow for more combinations and progressive variations.

6 Conclusion and Future Work

We presented an improved version of an emotional music generative system developed by Seiça et al. [5], with new parameters found in the literature as being relevant on music-emotions association. We used this system to perform an evaluation of these findings through a pilot user-test where most tendencies were confirmed. A few unexpected values were found, such as the low pitch in Anger, which was preferred by all the participants over the expected high pitch.

In future work, we expect to expand the musical features to enrich the musical scenario (e.g. timbre, rhythm) and perform an extended user test, with a larger sample of participants. The statistical analysis will also be detailed with pairwise comparisons to assess the significance of variations and sustained validation.

This pilot study was a first step to confirm general tendencies in emotion representation. These findings can be used to build an audio-visual computational artifact that evolves and generates outputs based on each individual's preferences, exploring the perceptual relevance on the visual domain and how it can be intertwined with the musical domain.

7 Acknowledgements

Mariana Seiça and Ana Rodrigues are funded by Fundação para a Ciência e Tecnologia (FCT), Portugal, under the grants SFRH/BD/138285/2018 and SFRH/BD/139775/2018. We also thank the twenty participants of our study.

References

1. Picard, R.W.: Affective computing. (1997)
2. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6) (1980) 1161
3. Brattico, E., Pearce, M.: The neuroaesthetics of music. Psychology of Aesthetics, Creativity, and the Arts **7**(1) (2013) 48
4. Gabrielsson, A., Lindström, E.: The role of structure in the musical expression of emotions. In: Handbook of Music and Emotion: Theory, Research, Applications. Oxford University Press, New York, NY, US (2010) 367–400
5. Seiça, M., Lopes, R.B., Martins, P., Cardoso, F.A.: Sonifying twitter's emotions through music. In: Music Technology with Swing - 13th International Symposium, CMMR 2017, Matosinhos, Portugal, September 25-28, 2017, Revised Selected Papers. (2017) 586–608
6. Edwards, M.: Algorithmic composition: Computational thinking in music. Communications of the ACM **54**(7) (July 2011) 58–67
7. Scirea, M., Togelius, J., Eklund, P., Risi, S.: Affective evolutionary music composition with metacompose. Genetic Programming and Evolvable Machines **18**(4) (2017) 433–465
8. Livingstone, S.R., Mühlberger, R., Brown, A.R., Loch, A.: Controlling musical emotionality: An affective computational architecture for influencing musical emotions. Digital Creativity **18**(1) (2007) 43–53
9. Software, K.: Generative music) (2018) `https://intermorphic.com/sseyo/koan/generativemusic1/`.
10. Eno, B.: Brian eno reflection (new ambient music) (2017) `https://brian-eno.net/reflection/index.html`.
11. Cope, D.: Experiments in musical intelligence `http://artsites.ucsc.edu/faculty/cope/experiments.htm`.
12. Juslin, P.N., Västfjäll, D.: Emotional responses to music: The need to consider underlying mechanisms. Behavioral and brain sciences **31**(5) (2008) 559–575
13. Juslin, P.N.: From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions. Physics of life reviews **10**(3) (2013) 235–266
14. Juslin, P.N., Laukka, P.: Improving emotional communication in music performance through cognitive feedback. Musicae Scientiae **4**(2) (2000) 151–183
15. Scherer, K.R., Oshinsky, J.S.: Cue utilization in emotion attribution from auditory stimuli. Motivation and emotion **1**(4) (1977) 331–346
16. Hevner, K.: Experimental studies of the elements of expression in music. American journal of Psychology **48**(2) (1936) 246–268
17. Balkwill, L.L., Thompson, W.F.: A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. Music perception: an interdisciplinary journal **17**(1) (1999) 43–64
18. Wedin, L.: A multidimensional study of perceptual-emotional qualities in music. Scandinavian journal of psychology **13**(1) (1972) 241–257
19. Cariani, P.: Music perception and cognition (2009)

Situating Research in Sound Art and Design: The Contextualization of Ecosound

Frank Pecquet,
Paris1 Pantheon Sorbonne University
fpecquet@univ-paris1.fr

Abstract. Discussed below are the implications resulting from the association of the following related disciplines: sound art and design. While each of these disciplines has a distinct meaning of its own, combining sound with art raises philosophical questions in aesthetics, even musicology, as sound in art tends to be thought of as a media to compose music. Combining sound and design may on the other hand lead to sound design, underlining a techno-scientific approach of a different nature linked to design sciences and techniques. Although this paper is not strictly dedicated to the comment on the notion of ecosound[1], it rather concerns general studies that may lead to contextualizing the notion itself, giving rise to epistemological questions. Moreover, this paper concerns "sound" in emergent sound disciplines, which may furthermore relate to the Anglo-Saxon concept of "sound studies" and which interests as well diverse fields of research, both techno-scientific and socio-scientific.

Keywords. Ecology; "ecosound"; listening; music; new sound order; sound design; sound art; design science, art work

Introduction

Mentioning the most recent changes in technology and ecology, I call for "a new sound order" [1] in which the neologism "ecosound" refers to all produced sounds in the post-industrial age and their mediation with regard to ecological awareness. While being ecological implies philosophical considerations about the post-traumatic stress in relation to "what we already know" [2] - global warming and the subsequent mass extinction -, design, as a social marker, indicates how cultural, political and economic changes have produced a "new sound order" in the Anthropocene[2]. The question as to whether or not we were already in a previous sound order logically comes to my mind as well. One may at the very least formulate this assumption - which by no means constitutes a voluntarist

[1] A term which means any artificial and/or functional sound conceived to be easily attuned with the environment or even improve it, be it a sound design to reduce sound in relation with the Anthropocene.

[2] Anthropocene is a more general term to describe geological changes produced on earth by men since they appear on the planet. Its sound equivalent would be the "anthropophony", an opposite term of "biophony" imagined by Bernie Krause in bioacoustics to describe all sound produced by living species except human.

proclamation but is induced by various works of research discussed below. For a long period on a human scale, all sound ecosystems are highly exposed to mass production - machines, tools, transportation, all are *sound facts* of human activities, and technology is the raw materials for this impulse.

While the main topic of this paper involves researches in humanities, I propose to reconsider the place of sound within the recently renamed "Art School of the Sorbonne University[3]" formerly called the visual art department. This paper was therefore written as a proposal to integrate sound studies in relationship with what is commonly defined as "research-creation" in art and design studies. To this extent, art doesn't refer explicitly to visual art, despite the natural tropism of the Sorbonne Art School towards this discipline, but to art in general, see the notion of art itself in relation to sound and music[4].

Design, as a scientific approach to creation, shares with art creative processes[5]. However, if art and design may be complimentary, design is always "functional" to comply with the user's requirements by definition. While the concept of creation qualifies the work of the final author, the notion of "planned creativity" [3] replaces it, to signify the creation without any artistic purpose, such as a production in the service of a good, a usage or information. Design products or services result from scientific approaches to conform with industrial specifications, while a work of art is not functional in essence and isn't conceived to be replicated. This serviceable attribute makes design creation exogenous to art[6]. Many designers conventionally refer to a design project as a three-fold methodology - analysis, creation, validation[7] -, it brings to the forefront aesthetic questions of substantial matter on the distinction of creative activities and objectives in art and design.

Finally, the expression "sound art and design[8]" represents a wide range of audible artistic activities that may be classified as sound art, but also any treatment and functional use of sound in non-artistic activities which is often the case in sound design[9]. Yet, the range of activities to be put at the account of sound art and sound design leads to a series of questions on the *medium* sound: does one have to consider that any mode of expression associated with sound can be legitimately classified in

[3] « *École des arts de la Sorbonne »* in which the research Institute ACTE (Art, Creation, Theory, Esthetics).

[4] "Sound" defines here the *medium*, while music is art.

[5] "Design is not an applied art, but the application of the idea of art in everyday life". Fabrice Bousteau (199, p1) in "What is design today" (in French) Beaux Arts Editions; Édition: Beaux-Arts Magazine (2009).

[6] Frank Pecquet. « The new sound order: Research and creation », HDR, Habilitation to Direct Research, Paris 1, Sorbonne University, 2018, IVb Chapter 3, "Sound design and creation: Conception and method" p.16.

[7] Nicolas Misdariis. « The science of sound design: Integrated approach of sound design within research in design », HDR, Habilitation to Direct Research, UTC Sorbonne University, 2018, Chapter 4, "Extension of the sound design domain", p. 59.

[8] Which stands for "sound art" and "sound design"

[9] I define sound design as "applied sound" in reference to the conventional name "applied art" for design studies. *Ibid*. IVb Chapter 2, "Sound design: Definition", p.4.

the category of sound art? And, if such is the case, how are various uses of sound likely to exceed the artistic dimension, a dimension legitimized by the expression "sound art[10]"? Likewise, doesn't sound design tend in the absolute to integrate what is today described as musical design [4]. All the more interesting questions exists for competences in specific fields of research - sound art and design. However, if these questions are not answerable at the musicological and philosophical levels, it's also because they mostly concern combined sectors of human activities that may put artistic issues in the background in favor of economic and cultural ones, covering societal needs, such as social housing, urban transportation, health, innovative actions, etc.

Sound Art and Music

The common classification of the artistic field gathered under the "musical genre" seems to persevere under that of "sound art", breaking away musicological tropism. In other words, as art may be common to both sound and music, art may borrow some of its concepts to music as well, be it non-artistic music, but sound anyway. There are other reasons to heighten musical judgment such as energy transmission for air molecule. Concrete music, electronic music, "acousmatic" music share pre-recorded sounds performed as electric current reconverted in organic sounds. Installation, sound poetry and sound sculpture are multimedia art where sound is juxtaposed and combined with other media. Whereas music is an artistic genre, the art to think and compose (with) sounds, sound art is a universal generic expression referring to all types of audio data used for artistic purpose.

If sound art may refer to music, it is not only limited to it. Sound may be associated with other media and, on one hand need be contextualized with other media, changing values by being heard in another context, for another raison. On the other hand, mixing media opens multi-dimensional perceptions by readdressing the process of listening in relation to different media, giving rise eventually to synesthesia. Consequently, sound art widens auditory perception to several senses but apart from any functional constrains. Sound art isn't limited to making or listening to music, but rather making or listening to different types of sounds, artistic or not, more concretely named as "environmental audio data". As such, it reveals the epistemological distinction of sound in general *and* the modes of expression and communication which characterize it in particular.

Art: Music versus Audio Culture

Sound as a media is common to both cultures, music and audio. However, the autonomous status of sound is proper to address the hypothetic distinction between audio and music culture which, nevertheless, remain interdependent. At the crossroad between art and design, there is an abounding development of new practices centered on sound which overall galvanizes functional and artistic

[10] Such expression would include in this case various sound production among sound installation, sound experiment, happening, field recording, poetry, theatre, noise, etc.

interrogations on audio culture - muzak, ambiental and/or environmental sounds, phono/sonotypes, audio advertisement/commercial music, audio branding, audio messages, audio signals and services, etc. All these specific sounds are used for specific functions, they are linked to products and services, without any artistic finality (although they borrow artistic features such as commercials using music). They are audio stream used in dedicated spaces in order to assist, alert, accompany Human in society.

Audio culture depends upon these phenomena as they impact the everyday life. Moreover, such practices are gradually expanding to this new sound order in which data banks filled with all possible sounds are available for everybody and vacant for all kinds of usage. While in this virtual audio tank, sound may be used for its own musical characteristics[11], they are available as shared audio material in big data center to valorize products and services. In such business context, sound becomes the audio exchange currency. Whereas music still remains the subsequent common denominator to such shared data, it is however subservient to artistic need in response to the global audio market in which such audio culture is bathing. Although the term music is universal, not all music may content consumer habits.

On the other hand, while art music[12] seems to remain well anchored in music culture (in a traditional sense), it still relates to concert music, whether it is classical or contemporary. As such, art music remains off the current from the audio culture previously described. It specially concerns higher educated audiences, reflecting institutional cultural tendencies. But art music has nevertheless evolved according to multicultural influences - as it has always been the case in music history - to encompass cultural differences in which popular factors have their parts. The notion of audio culture brings to light such distinctions. Audio culture allows a different acceptation of music (culture) when it comes to targeting functional needs, commercial or not. But such a change comes together with the technological revolution as well, from audio recording techniques (and audio broadcasting industry), digital signal processing and sampling techniques to AI and Big Data. All these changes in only a short century period. This explains the youth of audio culture and its close dependency upon new technologies.

Correspondingly, emphasis on sound has also evolved apart from the audio culture. The breaking of the tone and subsequent dissonance, serialism, avant-garde and experimental music led art music to experiment sound to such an extent that sound becomes the central issue of music today. It's like moving from a musical culture

[11] I do not include popular music, such as variety, pop songs, movie soundtracks, etc., since they are extra-artistic creation of commercial and/or functional orientation.

[12] The expression "art music" is used here to differentiate marketable musical productions without any artistic objective. Art music refers to written music from the occidental tradition, classical music is art music from a specific historical period for instance. This expression is used by German philosopher Harry Leymann in his book « The digital revolution in music », (2012, Germany, 2017, France). where he describes 1000 years of written practice in music composition up to the present time where computing samples questions the merits of writing music today.

centered on the tone to a sound culture. For contemporary composers, it means bartering your text book on tonal theory to that of sound technics. Writing music itself is questionable where machines may replace any human decisions with algorithmic procedures. To a certain extent, such changes lead to an interconnection between audio culture and art (music) favoring creative practices. Sound seems to have achieved its cultural revolution within this audio sphere. On one hand the media is recognized as free from aesthetic limits caused by a thousand years of canonical rules in art music and subsequent speculation. On the other hand, its status as the social marker of such audio culture reveals the acoustic organization of our daily audio environment.

Such transformations arouse auditive awareness while Humans have ruled the world by making it resemble them, as the most endangered species. The "anthropocentric equal temperament", as Timothy Morton writes, could be seen as a human obsession where everything else "becomes keyed to our teleological reference tone" [2 p.151]. According to Morton, Men have a tendency to annihilate the natural balance in between any species, see lifeforms by imposing their rule to conform to an artificial global agreement. Such view on this dampening effect to the biosphere echoed in Murray Schafer's book "Tuning of the world" [5] where he claimed, two generations ago, ecological awareness by taking inventory of all organic sounds to be attuned with. Nevertheless, if such considerations have resulted in concrete actions on reducing audio acoustic pollution - adjusting laws on noise regulation, audiometric controls, green pastille on household equipment, auditive health charter, etc., they remain symbolic and do not comply to ecological the collective awareness of the future. And sound is a matter of intensity, a paradigm for life: less power is less noisy, a possible slogan for the actual audio culture.

Design: Function versus Aesthetics

No need for further inquiry to attest that we live in an aesthetic epoch and that we are surrounded by works of art. Everyone is an artist or may view its own existence as an artistic challenge. Life itself is an aesthetic experience, a style, a design project. Life is perception were design is conception, the formal step in design is to model the idea, in any media. That is also what administration accomplishes in establishing cultural markers throughout concerts, exhibitions, art shows, etc. As stated by Gabriel Markus: "We, post-modern human beings, are witnessing a recent process of increasing aesthetics of our everyday objects" [6]. Another remarkable quote concerns Markus' intuition on art and design when he asserts that "it is more and more difficult today, in a society fully oriented towards aesthetics, to distinguish between art and design" [6 p.16]. If such ascertainment reinforces the idea according to which art has a direct aesthetic impact in design, one may ask further if art is the main function of design.

Talking about art and design, sound art and design more accurately, is also appraising the interaction between sound art and design therefore questioning the relationship between (sound) art and (sound) sound design. Art has multiple uses, it has no value in itself, but design can formalize art. The world of design is made on

consumption habits, but the dynamic function of design is not limited to services or use. The dynamic function of design is art, because art imposes formal beauty to ease consumers in their use of the product or services and please them. When art communicates beauty and pleasure, art fusions in design. A product (or a service) in design may be easy to understand or to play with, and beautiful to experience. The aesthetic function of design is to give the absolute power of imagination to any object or services with art.

If such statements motivate the aesthetic and functional quest, art influences design it is a fact, but design is the project. The "power of art" is the power of imagination to be attuned with the world (we design). Among various modes of sound manifestation to be considered in sound art - from noise production to visual art, sound art remains non-functional *per se*, qualifying music to be seen or touched within confined artistic boundaries. Sound design uses functional aesthetic to boost its usefulness.

Design research focuses then on the subsequent categories that we may associate with sound and design, see acoustic and design. In sound design, creation is tied to the logic of a design project [7]. Logically sound design is a sub category *sub-field* - of design. *Sub-fields* emerge with explicit audio concepts in sound studies: "everyday life"; "specific universe"; "marketing. They cover both ecological, geographical and interactive communication in a socio-technique environment. For the description: the *sub-fields* of sound design are in decreasing order after design as first field of the following ranking. All *sub-fields* are named according to what they do. Their name links to the sound design field as the second field after design: / Design // Sound_design /// Every_day_life //// Urban_environment. They are the chain link of four *sub-fields*: public_transportation\mobilities\working_spaces\architecture\education\profession\everyday_objects\natural_habitat_and_built_heritage\environment_and_care_equipment\...

Media: Sound Studies

The ambivalence of sound prevails in the definition of the sound itself. A sound is at the same time a physical vibration *and* the feeling which this vibration induces. The objective aspect of the sound is the natural object of sciences and techniques. Its subjective aspect depends upon the effect produced on the subject who feels it. The cause and effect relations work on three fields: physics, physiological and psychological. Listening to sounds is divided into several "listenings" in relation to the nature of the sound, but also to the listening itself, which is cultural. In this regard, the urbanist Philippe Amphoux's tripartition for listening hence described: memorized, reactivated and qualified [8] gives a fair representation of sound by the listener.[13] They respectively refer to memory, perception and interpretation.

[13] They can be compared to Truax' three Distinctions Listening cited [13 p. 143]: Listening - Research, Listening - Prediction and background Listening. Each of them can develop differently depending on the context.

These three forms of listening are represented by two pairs[14]: to listen and to understand, to perceive ("*ouir*") and to hear. The first pair indicates a process of listening that goes beyond the sound itself: to listen to someone does not only mean to listen to the sound of his voice, but what it describes in semantic terms. To listen allow defining sounds, identifying sounds in relation with the inner/outer environment. Sound has many life forms, they are waves between idea and object, but they are only real for the ears. The sound produces an effect in a sound design relationship: listening is being affected by the sound. The verb "to understand" shows an activity that is not necessarily linked to sound perception but to the reinforcement of what we designate here and there by *names*. In social acoustics, Patrick Susini's studies on the perception of everyday sounds classifies sound according to three different levels - *acoustics* concerning the perceived characteristics of the signal, *causal* which defines sound as a source index and/or the mode of production, and *semantics* that allows for the association of sounds with a same context (or "usage of scenario") [9]. Together with this audio-semantic perception and expression - to identify and to understand -, the pair - "to perceive ("*ouïr*") and to hear" -, describes in auditory perception a phenomenon of identification to qualify one sound amongst others. In all cases, sound is the *medium* and the *medium* is the message. In such equation, sound carries out more than mere aesthetic promises as well as services of common use, be it music, acoustic design, soundscape or lyrics, it provides meaning for human beings.

Ecosound: State-of-the-Art

The term "ecosound" is the contraction of two words: sound and ecology. I first used this term with regard to a project based on the rehabilitation of the Danish heritage canons, in this case the Gedser wind turbine[15], to work in favor of its artistic "rebound" against the backdrop of an ecological model. Initial research is found in class / Design// Sound design /// Every_day_life //// natural_habitat_and_built_heritage earlier described in design *sub-fields* at fourth level *sub-fields*[16]. But overall, the term "ecosound" is used in relation to global audio awareness. It concerns protecting acoustic environment from vibration to elaborate ecological norms, with the help of creative/innovative processes. If the ecosound allows measuring our audio tolerance to sound biodiversity, be it artificial, it also allows nurturing audio creativeness [10]. James Murphy's project "Subway Symphony" is an example of ecosound with creative intent. At the meeting of design and music, Subway Symphony can be considered as an archetype of urban sound design whose object is the improvement

[14] They are borrowed to Pierre Schaeffer's « Treatise of musical object » work published in 1966.

[15] The Gedser wind turbine is considered as the mother of modern wind turbine. For more infomation, see « L'éolienne de Gedser », Frank Pecquet in « Up Magazine » N° 3, 2018, and a web documentary at this address « http://www.ariadr.com/les-webdocs/gedser/ ».

[16] The practical research focuses on the development of a smartphone application interface for the management of sound production from wind energy, capturing and converting the data and formatting them as sound variables. In terms of creation, the tool makes possible, by selecting parameters, to recompose the energy as a sound continuum remodeled on the screen to produce music.

of subway sounds, especially the sounds emitted by digital turnstiles as a validation of the users' passage. The designer wants to remedy the current beeps described in Murphy's words as "flat and unpleasant" by substituting them for "softer" sounds, such as chimes judged and proven less stressful [11]. Initial research is found in class sound design> earlier described in design *sub-fields* at fourth level *sub-fields*/ Design // Sound design /// Every_day_life //// Transport.

The dynamics of sound art and design result, meanwhile, from various ways of the sound expression and production. The "multimodal sound[17]" not only proceeds from the organization of the sound parameters in the temporal scale, but it also proceeds from the combination of various modes of artistic trends - visual, plastic, theatrical, choreographic, architectural, cinematographic, and finally from design - computer, semiotic, decorative, beautiful, ecological. From the epistemological point of view, the terms "sound-noise", "sound body", "acoustic image", "musical object", "phonograph", "sonotype", etc., translate the fluctuation of statements to represent in essence a creative flow in this *new sound order*.

The study of various sound examples of our environment - largely debated for over 50 years [12], on "natural sounds" versus sounds produced by the Man and his machines, leads to scientific analysis of sound practices and uses. To measure the noise for regulation purposes, how and why a sound might affect auditory perception and influence a loss of listening may constitute one of the challenges to clarify among others. As such, it remains a basic step of the sound anthropology in this process of emergence which might require as well, as an adjustment with the environment, creative listening. Sound design puts it all together, there are numerous scientific publications on sound techniques from pedagogy to philosophy with an emphasis on fundamental acoustic, which is the nerd of experimentation in the three activity sectors of the ecosound system - physics, physiological and psychological. They correspond to specific sound studies.

The pair "perceiving and hearing" undertakes research on our *audible environment* in the context of design - and in parallel with that of sound arts -, to question how auditory processes might influence our decision to hear. At first, a sound conveys information (sound data). Second, listening to a sound is a cognitive process which allows identifying the source. Third evaluating a sound impact on the ear is like understanding its effect, when it is not a natural sound, to recognize and give the sound a name to make it useful, "designable". Within the stream of information, fragmented acoustic samples are concrete, "palpable"[18], immediate. Sound data opens acoustic communication, it directs listening. However, sound art and design translate sound data differently: if art expresses "sensitive sound" [13], design communicates by means of "graspable" and functional sound forms. One is implicit

[17] I use here the term « multimodal » in Y.Bellick's acceptance. The latter analyses the sound articulation with other modes of expression, such as the sound representation of external data in a navigable space (Yacine Bellick : *Interfaces multimodales : concepts, modèles et architectures* : (Doctorat thesis, University Paris-XI-LIMSI CNRS), 1995).

[18] [15 p.227, 260]

and aims at the "feeling of listening", the sound effect *per se*, the other is explicit and targets its object, sound data [14].

According to Jonathan Sternes "sound technologies are social artifacts from beginning to end" [15 p. 481]. Social artifacts are to be considered as global acoustic data. The global audio sphere has no real limit[19]. Experiencing audio is an awareness, whether it is a question of reaching new lifeforms. And experiencing audio is being aware of the acoustic environment through simple observation and/or analyze. Awareness is the result of a collective experience of the artifact with the *medium* sound.

To return to Sternes' view on social artifacts, technology is behind the process, as always. AI shows, to another extent, that being aware could also be artificial. It may lead to artificial knowledge, an ontological question. Machine's ecosystem is ruled by technological artifacts that materialize in communication as audio energy. The "Ecosonic artifact" refer to both ecological intent as consciousness, *and* audio data as sensorial knowledge. When sound "means" rather than "sounds", we are caught into a routine. Listening is like acting memory after the effect, there is nothing to hear or to listen. A meaningless sound is a sound that refers to an action we obsessively know as mental depression for example, a meaningful sound could be the issue. That is also why making a sound awakens brain, historical awareness is emotional intelligence. In semantic, sound is a (sound) idea, an "acoustic image". It is materialized in large audio banks where each *ready-to-sound* vibration is catalogued to answer a specific search. The internaut puts the "vibe" in context by classifying data types in large audio banks, accessible by keyword tracking. This "open access" Meta Data party, see Mega Audio Data, is another data generation of AI working on retrieving information to compose "virtual audio matter".

The formal distinction of sound art and sound design is a matter of finality. Sound art puts existential questions on making an object, which first is meant to be actioned (activated) by a user. There is no question on how to use it or what it is. It is a work of art with its own sound ecosystem. Sound design concretizes the idea by formally adapting its representation in a dedicated object [16]. The usage and related services generate the noises of usage and the sounds of services, they lead to a concrete listening, calling for the process of identity and its consequent analysis. In this context, sound design maintains a specific relation with its object. The sound is apprehended as "meaningful", the listening is associated with a palpable function - to alert, inform, accompany, arrange, adapt: the produced effect causes a reaction, involves a state, instigates an awakening. As a reproducible industrial artifact, the sound object (attribute, property, data bank) in sound design sciences is a sensory marker assisting hearing in the audible environment. The latter leads to a better adaptation of acoustic products and services which accompany Men in society

[19] Because silent is audio only if we think audio, it is reached when there is no work of any kind in the atmosphere, no tension, no physical crisscrossing, just "no sound". This is our sound operating system. Audio data are like sound waves intertwined within a larger lifeform system.

among their various activities. The question of regulating such a stream needs be ecological, therefore political.

Sound Facts and Listening Contexts

Common denominator of art and design, the sound ("*le sonore*") as well qualifies the sounds *in situ*, i.e. sounds directly drawn from the environment. Associated with the effects generated by the various sound manifestations, "raw sound" is distinguished from the musical sound. One and the other answer specific modes of auditory awareness - to perceive, to listen and to hear. In the auditory perception, the "musical" is not opposed to the "sounding", it differentiates a state in motion from a banal listening towards an enchanted listening of sounds. The music of sound[20] qualifies the process of subjectivation occurring when hearing in accordance with the articulation of the above mentioned three faculties: perception, listening and hearing, as intertwined processes. The sound is first a physical phenomenon, listening is a psycho-physiological process. Research in sound design consequently also concerns auditory phenomena - from perception and hearing -, fundamental sciences both apply in physics and social sciences- listening situations (*in-situ*), in a way to analyze sound integration according to the different listening types of practices [17][21].

Theoretical research on sound design thus leads to social sciences giving to sound production a social dimension: interior architecture and equipment, transport and place of work, social space and urban environment... Although the sound remains a channel of communication naturalness, in human representation, the sound infiltrates all eco sectors. To increase harmony between man and his direct environment, the sound ecology arbitrates the noise tolerance in the auditory biosphere [18]. But the sound ecology is also concerned with the regulation of the sound inheritance in extinction, among which animal species extinct and global warming facts, typical anthropological data such as forgotten ceremonies, past practices and techniques, lost civilizations, outmoded tools, out of cycle life machines, cultural traditions, etc.

In the deontology of listening, one must know how to distinguish which listening to adapt in relation to which ear. Sound is inseparable from temporality from which listening manifests. Considering the relationship of sound with the perceiving subject involves philosophical questions about formal intentions in sound conceptualization. How does sound make us become aware of reality? How does sound influences things around us? How it affects existence? World is full of sound phenomena that produce streams of audio data [19]. A world completely quiet, anechoic, remains purely utopian, and artificial. A universe without any phase, neither period, nor frequency appears even more speculative. Listening is a reactive process, it requires time shifting appreciation. But listening is also feeling behind a

[20] But also, the sound of the music, to borrow a metaphor of "acousmatic music".
[21] As *sub-fields* earlier described in the methodology.

conceptual experience[22]. Phenomena are sound representations. They may lead to "musical facts" through reasoning.

M. Schäfer in his reference work "The Soundscape: Our Sonic Environment and the Tuning of the World" recommends using the expression "sound fact" rather than "sound object" for the sake of sound analysis in context or within an environment such as signals, symbols, tonalities or prints. He writes [11 p.195]: "The *sound fact* distinguishes from the sound objects, specimens of laboratory. The *sound fact*, by the definition which the dictionary of the fact gives, namely *what arrived, which takes place*, better suggests the idea of an existence compared to a context. Thus, the same sound of church bells, for example, can be regarded as a sound object if it is recorded and analyzed in a laboratory, or as a sound fact if it is identified and studied within the community. The studies on the sound fact distinguish three criteria: physics: acoustic contents; reference frame: origins and indices; aesthetics features[23]."

Perspective in Sound Design

Contemporary philosophy of design claims humanism, social and environmental well-being, and makes design a noble discipline [20]. The science of sound design, as a sub-category of design, therefore, respects such ideals. According to the design philosophy, notions such as living space and/or acoustic quality share common dynamics that would as well contribute to enhance listening awareness, making more pleasant the sound environment and/or giving a better appreciation of the sound of nature. More generally, sound optimization means being ecological, moreover being "ecosound" responsible. From the standpoint of design, the sound is seen as an object of life, a living phenomenon with contextual aspects such as functional characteristics - dynamic effect, space identity -, historical sources, aesthetic territory naturalness, all types determining specific auditory ecosystems. In a limited space-time perimeter, aesthetic research studies the relationship that may occur between sound and music, stylistic tendencies of listening territories, cultural acoustic rituals issued from sound craft, industrial and art work. Designing sound is both a creative activity when it is about producing appropriate sounds and the adjustment of the existing ones. As the architect thinks space according to the inhabitant mobility and activities, by enhancing a specific neighborhood, the sound designer improves the audio ecosystem, everywhere he thinks that acoustical environment needs and allows it: not inevitably by producing new sound sources but by improving the relationship between sound and people.

Sound design also concerns the use of sounds and their possible manipulation, as an instrument to guide you in action, operated in relation to all types of community, the sound of an isolated individual, a group in a virtual audio network. Sound anthropology studies sounds within scales - temporal, socio-economic or aesthetic -,

[22] Emmanuel Kant.
[23] [17] p.195.

according to various contexts, in various interactions to organize a sound ecosystem: inspecting sound levels could be viewed as protecting health - vocal, instrumental, environmental sources -, in relation to human activities in public/private spaces. Sociology of sound studies the sound stream within an ecosystem, in a given context, by confronting various sounds in various spaces: districts, markets, places of worship, types of transport, at home or on the place of work, to intervene in the development of the sound charters - diagram, graphs, charts, sites of the place with various parameter settings: on a side confined sound space, but also other environment such as not-controlled spaces. The science of sound design works on multi-purpose sound usage - *factual sounds* - and tends to ease the relation between sounds and what we hear, the audible limit in this relation, after analyzing how they sound [21].

Conclusion

Taking into account the various aspects of sound art and design leads to some general observations. First, considering sound in relation to art leads to the creative aspect of sound production, which one involves all sounds and all consequent hearing experiences. This involves aesthetic motivations more specific to artistic expression, and aims at producing a work of art. Second the functional aspect of the sound design project, whether or not it integrates aesthetic considerations such as "beautiful" and "agreeable", answers industrial (reproducible) constraints to match functions with ideas (alert, memory, identity). Third, if art such as "bio-art" integrates ethical consideration with regard to ecology, its premiere motivation is to reveal a state of mind from a specific artwork by opening individual awareness to reality.

As a medium, if sound is always the result of hearing, external ear formalizes audio codes. Sound art and sound design are distinct field with probable *sub-fields* connected[24]. While the "ecosound" in the sound design context is meant to focus on being ecological in producing a sound useful to in acoustic with an emphasis on environmental awareness, whatever it is applied to: object/services, message and spaces, the science of design works out a protocol of specific rules for sound optimization in sound design: controlling sound production, renewing audio system protocols, humanizing all audible communication networks, objects and services [22]. Among these daily activities, it is thus necessary to manage the space time of the sound environment, to accompany sound perception on a time progress scale, to familiarize listening and create promising acoustic ecosystems, what would best define the "ecosound" as the unit of the "new sound order", building up a sound design inventory of sound designers, studying practices, products, services, unearth artifacts.

[24] When a sound is "plastic", it is a visual art piece, whatever it sounds like.

References

1 Foucault, Michel. L'ordre du discours (in French. trad. “The Order of Discourse”), Paris, Éditions: NRF, Gallimard, (1970).

2 Morton, Timothy. Being Ecological. A Pelican Book, Penguin Random House, UK, 2018.

3 Archer, Bruce. Connaissance du design: la créativité planifiée en industrie. “Design awareness and planned creativity in industry”, London, Design Council of Great Brittain (1974).

4 Boumendil, Michael. Design musical et stratégie de marque. Quand une identité sonore fait la différence! “When a Sound Identity Makes a Difference” (in French), Paris, Éditions: Eyrolles, (2017).

5 Pecquet, Frank. Thèse d’Habilitation à Diriger les Recherches, Vol. IVb Recherche en design sonore. (in French trad. “Research in sound design”), Université Paris1 Panthéon-Sorbonne, (2018).

6 Gabriel, Markus. Le pouvoir de l’art. “The Power of Art” (in French), Éditions Saint Simon, Paris, (2018).

7 Pecquet, Frank, Misdariis, Nicolas, Donin, Nicolas, Zattra, Laura, Fierro, David. Analysis of Sound Design Practice (ASDP): Research Methodology. In : XXII CIM (Colloquio di Informatica Musicale), Udine, Italie, (2018).

8 Amphoux, Pascal. L’identité sonore urbaine : Une approche méthodologique croisée). “Urban Sound Identity: A Cross-Methodological Approach”, (in French) Grenoble, Publication PDF par le CRESSON (Centre de Recherche sur l’Espace Sonore et l’Environnement Urbain), (2003).

9 Susini, Patrick. Le design sonore : un cadre expérimental et applicatif pour explorer la perception sonore. “Sound Design: An Experimental Framework to Explore Sound Perception”. (in French), Thèse d’Habilitation à Diriger les Recherches, (2011).

10 Pecquet, Frank *Echo Design.* In : “*The 22nd International Conference on Auditory Display* (ICAD-2016) July 2-8, 2016”, Canberra, Australia, (2016).

11 Sinclair, Peter. « Vivre avec les alarmes : l’environnement sonore d’un service de réanimation et soins intensifs. “Living with Alarms: the Sound Environment of an Intensive Care Unit.”, (in French) dans Locus Sonus : 10 ans d’expérimentations en arts sonores. Éditions: Le mot et le reste, Coll. Carte Blanche, (2015).

12 Schäfer, Murray. Paysage sonore : le monde comme musique. “The Tuning of the World” (in French), Marseille, Éditions: Wild Project, (2010) [1977].

13 Zourabichvili, François. L’art comme jeu. “Art as Game” (in French) Préface de Jean Luc Nancy. Presse Universitaire de Nanterre, (2018).

14 Candau, Joël, Le Gonidec, Marie Barbara. Paysages sensoriels: essai d'anthropologie de la construction et de la perception de l'environnement sonore. “Sensory Landscapes: Essays on Anthropology for the Construction and Perception of the Sound Environment”, (in French) Paris, Éditions: Comité des travaux Historiques et Scientifiques, (2013).

15 Sterne, Jonathan. Histoire de la modernité sonore “The Audible Past: Cultural Origins of Sound Reproduction” (in French), Paris: La Découverte, (2015) [2003].

16 Marry, Solène (2013). L'espace sonore en milieu urbain. “The Sound Space in Urban Environment”, (in French) Rennes, Éditions: Presses Universitaires de Rennes.

17 Guiu, Claire, Faburel, Guillaume, Mervant-Roux, Marie-Madeleine, Torgue, Henri, Woloszyn, Philippe. Soundspace : espaces, expériences et politiques du sonore. “Soundspace: Spaces, Experiences and Sound Policies”, (in French) “Rennes, Éditions: PUF Rennes, (2015).

18 Mariétan, Pierre (2005). L'environnement sonore : approche sensible, concepts, modes de représentation. “Sound Environment: Sensitive Approach, Concepts, Modes of Representation”, (in French) Paris, Éditions: Champ Social Édition UNESCO.

19 Findeli, Alain. In Sciences du design, Vol.1. “La recherche-projet en design et la question de la question de recherche: essai de clarification conceptuelle”. “The Research-project in design and the question of the research question: conceptual clarifications”, (in French) PUF, Paris, (2015).

20 AFD. Alliance Française des Designers, http://www.alliance-francaise-des-designers.org/ consulté le 10 mars 2015.

21 Vial, Stéphane. Le design. Paris, Éditions : Que sais-je ? (2015).

The Statistical Properties of Tonal and Atonal Music Pieces

Karolina Martinson[1,2] and Piotr Zieliński[1,3] ⋆

[1] Institute of Nuclear Physics Polish Academy of Sciences
[2] AGH University of Science and Technology in Cracow
[3] Cracow University of Technology
karolina.martinson@ifj.edu.pl

Abstract. Two solo flute pieces: *Allemande* from *Partita in A minor*, BWV 1013 by Johann Sebastian Bach and *Syrinx* L. 129 by Claude Debussy, were analyzed. The article presents the results obtained from the statistical analysis of ascending and descending intervals, pitches (frequencies of the sound) and duration times. It was found that in the double logarithmic scale histograms of occurrence frequencies exhibit significant deviations from linear functions. The gradient of a line were usually smaller than -1, which indicates that Zipf's law is not fulfilled. Autocorrelation functions and periodograms have been calculated. The aim of this article is to compare pieces that are tonal and atonal. The frequency histograms allow one to classify the time signals as 'colourful noises'. *Syrinx* is close to the pink noise and *Partita in A minor* to the brownian noise.

Keywords: atonal and tonal music, Syrinx C. Debussy, Partita in A minor J. S. Bach, Zipf law, statistical analysis

1 Introduction

Mathematical and physical properties of time series as are musical pieces can be examined using statistical methods. The existence of such principles in the products of human creativity was discovered at the beginning of the 20th century. Jean-Baptiste Estoup in his work [1] described the occurence frequency of words in wtirren texts and noticed that it fulfills the rule called later Zipf's law. Similar researches were carried out since 1970s for musical works. Results obtained by various authors show that Zipf's law is fulfilled for music.

The autors [2] found that the pitch and loudness fluctuation in classical, jazz, blues and rock music follow Zipf's law. It was figured out that the various values related to the progress of pieces have a power spectra close to $\frac{1}{f}$ noise - pink noise. In the article [3] changes with the frequency of the Bach's pieces fulfilled Zipf's law. Manaris, Purewal and McCormick [4] examined different authors and genres. A pitch and duration time also fulfilled the law [5][6]. Distinction between

⋆ KM has been partly supported by the EU Project POWR.03.02.00-00-I004/16.

ages or authors can be possible using numerical analysis [7][4] however, the analysis itself is not sufficient enough to say whether the music is *pleasant*. The definition of the *pleasant* word is not simple. Every person defines it in a different way. Psychologists, philosophers, art theorists etc. have been trying to devise a definition of *pleasant* and *beauty*. Mayer [8] suggests that the emotional states in music (sadness, anger, happiness) are defined using statistical parameters i.e dynamics, octaves/registers, continuity, speed. Although these parameters vary locally within a progress in music, they are relatively constant globally [4]. If the frequency of appearing elements in an analyzed phenomena shown in the double logarithmic scale gives a straight line with a slope equal to -1.0 the Zipf's law is fulfilled. Research proves that this is certainly true for literature [9] and music [2]. Two pieces for solo flute were analyzed:

1. *Syrinx for solo flute* - C. Debussy is a piece of music for solo flute written in 1913. Harmonic functions (chords) are not clear. Debussy composed in impressionism. The main component of this period is colour, chords are timbre spots. Debussy composed without regarding rules of tonal harmony. Tonal relation between two chords/sounds do not exist in his piece *Syrinx*. Dissonances which introduce harmonic tension and prefer to 'resolution' lose their current role. All kinds of scales i.e pentatonic and full-tone scale were used by the author. The contrast of registers (low and high pitches), complicated rhythmic figures (quintolets or sixteenth triplets and octal triols), diverse and contrast dynamic are also visible in *Syrinx* [11].
2. *Allemande from Partita in A minor for solo flute* - J. S. Bach, BWV 1013, wrote probably in 1723. The suite consists of four movements. The first movement is Allemande - 'german dance', a renaissance and baroque dance in duple metre often consisting of only sixteenths notes. Harmonic functions (chords) are clear [10].

2 Methods

To check how many times a given pitch or rhythm are repeated it was necessary to use computer software which counts amount of repeated intervals (distance between two sounds) and rythmic values (duration time of the sound). Zipf's law is fulfilled when the slope α in the double logarithmic scale plot is equal to -1. The results of calculations can also adopt other forms. The deviation from the straight line shape describes the coefficient of determination R^2 [12]. This parameter can take values from 0 (wrong fitting) to 1 (satisfactory fitting). It evaluates the quality of the model's fit.

Determination coefficient R^2 expresses the proportions between the points variability represented by the trend line [13]. This value is calculated automatically in *Origin* software, as a fitting function using the following equation:

$$R^2 = \frac{\sum_{t=1}^{n}(\widehat{y}_t - \overline{y})^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}, \tag{1}$$

where y_t is the actual value of the variable Y in time t, $\widehat{y}_t$ is the theoretical value of variable explained in the model, $\overline{y}$ is an arithmetical mean.

3 Results

3.1 Partita in A minor, J. S. Bach

Intervals Figure 1 shows log-log function including frequency of occurring ascending and descending intervals. α coefficient is close to -2. Determination coefficient is equal to 0.82.

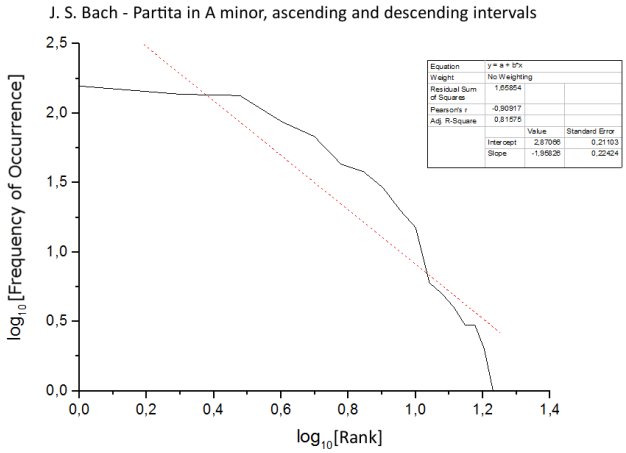


Fig. 1: Rank-Frequency distribution of intervals in *Allemande* from *Partita in A minor*, J. S. Bach, $y = -1.96x + 2.87, R^2 = 0.82$

Pitch Zipf's law has been reached for the pitches, α - slope value, is close to -1 [Fig.2].

Rhythm The sixteenth notes are prevail in Bach's piece analysed here. The slope is high value equal to -4.19 [Fig. 3].

3.2 Syrinx, C. Debussy

Intervals Figure 4 shows log-log function including frequency of occurring ascending and descending intervals. α coefficient is close to -2.04. Determination coefficient is equal to 0.9. *Syrinx* as an atonal piece has a majority of small intervals. One semitone (minor second) and two semitones (major second, whole ton) are intervals which are the most populated in this piece. Zipf's law has not been satisfied because of α closes to -2 [Fig. 4].

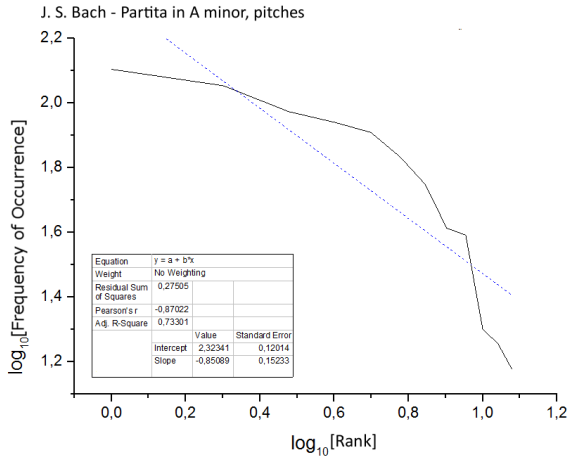


Fig. 2: Rank-Frequency distribution of pitches in *Allemande* from *Partita in A minor*, J. S. Bach, $y = -0.65x + 2.32, R^2 = 0.73$

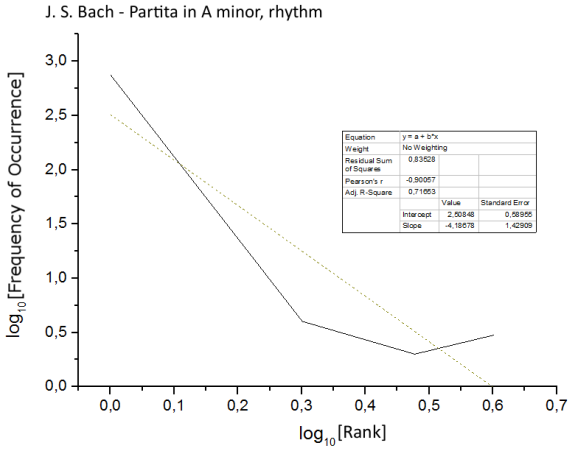


Fig. 3: Rank-Frequency distribution of rhythm in *Allemande* from *Partita in A minor*, J. S. Bach, $y = -4.19x + 2.51, R^2 = 0.72$

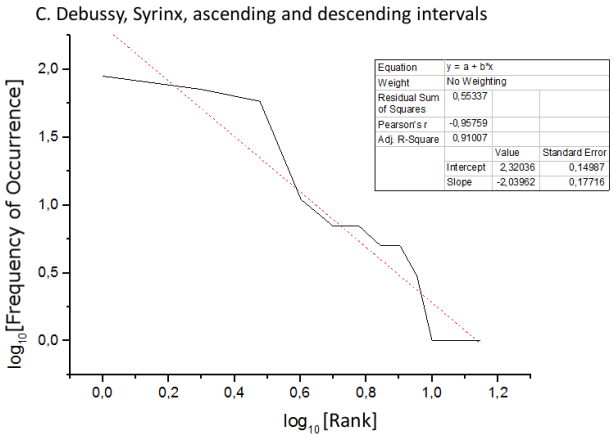


Fig. 4: Rank-Frequency distribution of intervals in *Syrinx*, C. Debussy, $y = -2.04x + 2.32, R^2 = 0.91$

Pitch Every pitch was analyzed without taking into account the octave of the pitch. The most popular pitch is *d-flat*.

Rhythm *Syrinx* is more complicated piece than *Allemande* as far as the rhythm is concerned. Demisemiquavers, sixteenth triplets and sixteenths are most common rhythmic values in *Syrinx*.

4 Discussion

4.1 Allemande, Partita in A minor, J. S. Bach

A fit to a single straight line is visibly not adequate. For a better analysis it was necessary to use two fitting lines. The first line for point's range from 1 to 3, $\alpha = -0.15, R^2 = 0.88$, and the second for points from 4 to 23 range $\alpha = -3.15, R^2 = 0.94$, which do not meet the requirements of Zipf's law.

Only for ascending intervals the α coefficient is equal to -1.78 and for descending intervals α is equal to -1.91 both cases seem to be Brownian noise. For sounds with two fitting lines, points from 1 to 9 $\alpha = -0.53$ and $R^2 = 0.81$, for points 9 to 12 $\alpha = -3.14$ and $R^2 = 0.81$. For satisfying R^2 black noise is obtained and Zipf's law is not fulfilled. Considering all sounds with distinction of octaves the most populated sound is e^2 (2 line octave) as a middle register of the flute. E sound can be fifth of the tonic chord or prime of the dominant chord in A minor. $\alpha = -1.17$ so it can be seen that Zipf's law is almost satisfied. Two fitting lines for points 1 to 8, $\alpha = -0.33, R^2 = 0.94$: white noise, for points $8 - 32$, $\alpha = -2.39, R^2 = 0.83$: brownian noise. Coefficients values are different. Results do not meet the requirements of Zipf's law. Zipf's law is fulfilled for one

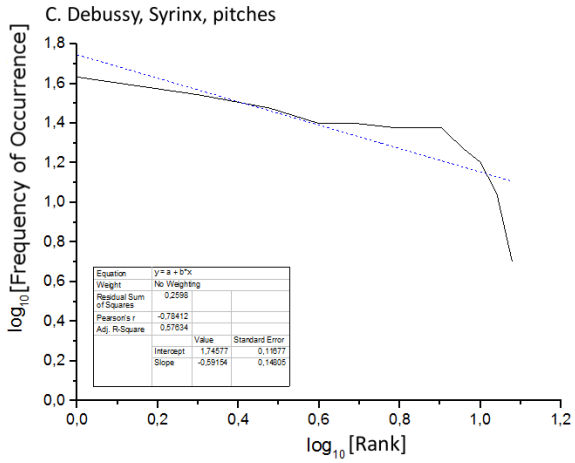


Fig. 5: Rank-Frequency distribution of pitches in *Syrinx*, C. Debussy, $y = -0.6x + 1.75, R^2 = 0.58$

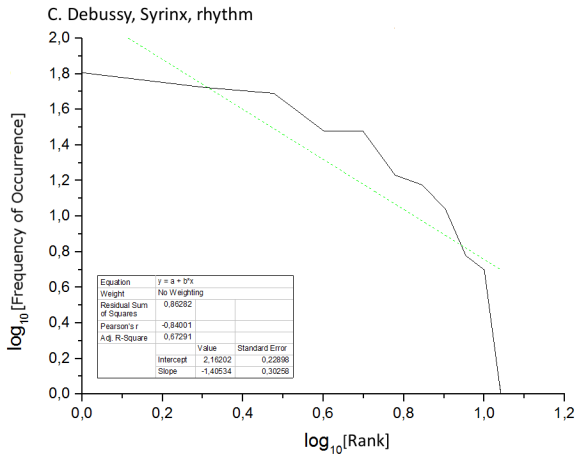


Fig. 6: Rank-Frequency distribution of rhythm in *Syrinx*, C. Debussy, $y = -1.41x + 2.16, R^2 = 0.67$

fitting line in sounds analysis. When fitting is improved (two fitting lines) Zipf's law is not fulfilled but R^2 is significantly better. Brownian noise is produced for ascending and descending intervals.

4.2 Syrinx, C. Debussy

The most populated interval is major second for ascending intervals and minor second for descending intervals. Small intervals are the most common. For ascending intervals (one fitting line) $\alpha = -1.69, R^2 = 0.91$ and for descending intervals (one fitting line) $\alpha = -2.12, R^2 = 0.90$: brownian noise in both cases. For two fitting lines: points from 1 to 3 range it is obtained $\alpha = -0.91, R^2 = 0.98$: Zipf's law fulfilment; points from 3 to 15 it was obtained $\alpha = -2.87, R^2 = 0.85$: black noise.

The most common sound in *Syrinx* part is sound *d-flat*. The key of the pieces is D-flat major and the part is ending on the d-flat sound. Second frequent sound is *b-flat* so it is third of the subdominant chord. Despite the presuption that *Syrinx* is atonal piece, statistics tells us that the tonality exist. The lack of the scale is heard but is not visible from the statistical point of view. One fitting line shows $\alpha = -0.59, R^2 = 0.58$. Two fitting lines for points 1 to 6 $\alpha = -0.36, R^2 = 0.94$, from points 7 to 12 $\alpha = -2.48, R^2 = 0.73$. Fitting functions are better but Zipf's law is fulfilled for the first point's range. It seems to be Brownian noise for the next 6 points. The sounds with their octaves were analysed here. The most frequent sound was b-flat (one line octave), $\alpha = -0.94, R^2 = 0.62$: Zipf's law is fulfilled.

5 Conclusions

Allemande from *Partita in A minor* J. S. Bach and *Syrinx* C. Debussy were analysed in this article. Zipf's law is not fulfilled for the intervals in both pieces. Slope coefficients (α) are close to -2 which is not an expected value (-1). Thus, the organization of sounds in these pieces resemble Brownian noise. Zipf's law is better fulfilled for the pitches (the frequency od the sound). Slope coefficient is closer to -1 but determination coefficients are worse ($-0.58 - 0.73$). Slope coefficients in rhythm in both pieces are different. In *Syrinx* is equal to -1.14 (Zipf's law is fulfilled), but in *Allemande* is equal to -4.19 (black noise). The authors [9] claimed that the characteristic of texts fulfilled Zipf's law with similar slope coefficients, however it seems not to be realized in these pieces. α has different value in rhythm, intervals and sounds (pitches) analysis. One line fitting is not enough to describe the slope coefficients. Few line fitting gives better quality but Zipf's law is not fulfilled. The analyzed pieces are different from the musical point of view but from this statistical point are similar. It is possible that the monophony of the pieces contributes to the similarity. Monophonic pieces have to interest the audience despite the limited elements of musical expression i.e. polyphony, harmony, different timbre or notes length. Bach wrote *Allemande* using almost sixteenth notes.The flute which was used in baroque differ from the

flute for which Debussy wrote in impressionism. Debussy could allow himself to write more technically complicated pieces [14][15].

References

1. Jean-Baptiste Estoup. Gammes stnographiques 3d ed. 1912
2. Voss, R. F., Clarke, J.: '1/f noise' in music: Music from 1/f noise. J. Acoust. Soc. Am. 63, 258 (1978)
3. Hsu, K. J., Hsu, A..: Self-similarity of the '1/f noise' called music. Proc. Nati. Acad. Sci. USA 88, 3507-3509, (1991)
4. Manaris, B., Vaughan, D., Wagner, C., Romero, J., Davis, R. B.: Evolutionary Music and the Zipf-Mandelbrot Law: Developing Fitness Functions for Pleasant Music. In: Cagnoni S. et al. (eds) Applications of Evolutionary Computing. Lecture Notes in Computer Science, vol 2611. (2003)
5. Zanette, D. H.: Zipfs law and the creation of musical context. Consejo Nacional Investigaciones Cientificas y Tecnicas Instituto Balsiero, Argentina 2008.
6. Rafailidis, D., Manolopoulos, Y.: The Power of Music: Searching for Power-Laws in Symbolic Musical Data. Department of Informatics, Aristotle University, Greece.
7. Manaris, B., Juan R., Machado, P.: Zipfs Law, Music Classification. In Comput. Music J. 29/1, 2005
8. Meyer, L.B.: Music and Emotion: Distinctions and Uncertainties. In: Music and Emotion Theory and Research, Juslin, P.N., Sloboda, J.A. (eds), Oxford University Press, Oxford 2001.
9. Grabska-Gradziska, I., Kulig, A., Kwapie, J., Owicimka P., Drod, S.: Multifractal analysis of sentence lengths in English literary texts. AWERProcedia Information Technology & Computer Science 03, 1700, 2013
10. Partita in a minor, J. S. Bach `https://en.wikipedia.org/wiki/Partita_in_A_minor_for_solo_flute_(Bach)`
11. Syrinx, C. Debussy `https://en.wikipedia.org/wiki/Syrinx_(Debussy)`
12. Wtroba, J.: Prosto o dopasowaniu prostych, czyli analiza regresji liniowej w praktyce. StatSoft Polska Sp. z o. o. 2011.
13. Coefficient of Determination `https://www.britannica.com/science/coefficient-of-determination`
14. Wolfe, J., Smith, J., Fletcher, N., McGee, T.: The Baroque and Classical Flutes and The Boehm Revolution. In: Bonsi, D., Gonzalez, D., Stanzial, D.: Proc. International Symposium on Musical Acoustics, Perugia (2001).
15. Wolfe, J., Smith, J., Tann, J., Fletcher, N.H.: Acoustic impedance of classical and modern flutes. J. Sound &Vibration, 243, 127-144, 2001.

Webmapper: A Tool for Visualizing and Manipulating Mappings in Digital Musical Instruments

Johnty Wang[1], Joseph Malloch[2], Stephen Sinclair[3], Jonathan Wilansky, Aaron Krajeski[1], and Marcelo M. Wanderley[1]

[1] Input Devices and Music Interaction Laboratory, CIRMMT, McGill University
{johnty.wang, aaron.krajeski}@mail.mcgill.ca,
jonathan.wilansky@gmail.com, marcelo.wanderley@mcgill.ca
[2] Graphics and Experiential Media Lab, Dalhousie University
joseph.malloch@dal.ca
[3] Multimodal Simulation Lab, Universidad Rey Juan Carlos
stephen.sinclair@urjc.es

Abstract. This paper describes the motivation, implementation, and usage of the application *Webmapper*, a tool for visualizing and manipulating mappings in the context of digital musical instrument (DMI) design. *Webmapper* is a user interface for interacting with devices on the *libmapper* network, a distributed system for making dynamic connections between signals within discrete devices that constitute a DMI. This decoupling of the mapping as a separate entity allows flexible representation and manipulation by any tool residing on the network—exemplified by *Webmapper*. We demonstrate the capability and potential utility of providing different representations of mappings in the work-flow of DMI design under a variety of collaborative and individual use cases, and present four visualizations applied to mappings used from a previous project as a concrete example.

Keywords: mapping, DMI design, prototyping, collaboration, visualization

1 Introduction

Mapping, in the context of digital musical instruments (DMIs)[1], pertains to the translation of input signals into resultant sound. Since mapping determines the ultimate behaviour of the instrument, it is an important part of the design process and an interesting area of research [2].

This paper describes the motivation, implementation, and usage of the application, *Webmapper*, a tool that supports multiple approaches to visualizing and manipulating mappings in the context of DMI design. First, the contexts which inspired *Webmapper* are introduced including a brief introduction to the *libmapper* framework that provides the underlying connectivity features. Then, the structure and implementation of *Webmapper* is presented using examples of

mappings from projects demonstrating the various visual representations implemented. Finally, an evaluation of the different views are presented, along with a discussion and future work related to the application.

2 Background and Related Work

2.1 Mapping Tools for DMI Design

A number of general-purpose graphical tools, designed for building interactive and multimedia systems, are used by the community to create mappings in the context of DMI design. Some—such as Max[4], Pd[5], and TouchDesigner[6]—provide full programming environments that can be used to define the structure of the entire instrument. These tools not only allow visual representation of the connection and processing of signals that make up part of the mapping process, but also can embed the interfaces to hardware and software components related to the sensor input devices as well as output synthesis systems. In terms of representation, these visual environments provide a signal-flow or "patching" interface that resembles the physical connection of wires in an audio processing chain.

There are also toolboxes and applications dedicated to the process of designing mappings specifically for DMIs. Some examples such as OSCulator[7] and junXion[8], are standalone applications that provide drivers for hardware input devices and allow transmission of signal data to programmed endpoints with data scaling. Other toolboxes provide specific mapping features to an existing environment such as a library of mapping and signal conditioning primitives [3], matrix-based manipulations specifically for mapping [4], mapping between different dimension spaces via geometric representations [5], or creation of mappings via machine learning [6].

One key commonality among all these existing tools is that they provide a single method of representing the mapping. *Webmapper*, on the other hand, allows more than a single way of representing and manipulating the mapping structure. The Jack Audio Connection Kit [9] provides an API that allows applications to access and modify the audio and MIDI connections between virtual endpoints on a local system, which results in the possibility of multiple command-line and GUI tools. However, Jack was designed to work with connections only.

2.2 libmapper

Through working on a number of collaborative projects involving DMIs spanning more than 10 years, a software framework for creating dynamic mappings,

[4] `https://cycling74.com/products/max/`
[5] `https://puredata.info/`
[6] `https://www.derivative.ca/`
[7] `https://osculator.net/`
[8] `http://steim.org/product/junxion/`
[9] `http://jackaudio.org/`

libmapper [7], was developed. Some concepts that prompted the development of *libmapper* include:

- **Experimentation:** The design of DMIs involves many variables such as the selection of sensing components, mapping, and synthesis techniques. These are not standard procedures and the process often involves exploration and experimentation.
- **Diversity:** Since work with DMIs often involves collaborators from different backgrounds, there isn't a single tool or approach that will work for everyone. Fixed representation standards may be limiting.
- **Distributed control:** Under collaborative contexts, it may be useful to allow multiple users to view and modify the mapping configuration at the same time.

To facilitate experimentation, it is necessary to provide the ability for connections between components to be quickly created and modified. The diversity of users suggest it may be useful to provide more than a single view and interaction method on the state of the system. Concurrency implies the need for a network based model where more than a single user can access and manipulate data at the same time. As a result, *libmapper* was developed as a framework upon which more modular and flexible approaches to mapping design can be realized. At its core, *libmapper* as a software library provides the means to expose a device to a network that allows automated discovery and dynamic connection with signal conditioning built into the connection itself.

The fundamental components on the *libmapper* network are *devices* and *signals*. An input device may contain a number of output signals which may for example be values corresponding to sensors intended to measure a set of gestures, and an output device such as a synthesizer will feature input signals that correspond to control parameters that affect the generated audio. *libmapper* allows *links* to be made between devices which construct high level associations between devices, and *maps* which are dataflow connections between parameters of interest. Another key feature of *libmapper* is that some basic signal conditioning can be built into the connection itself so that commonly used methods such as scaling, clamping, and basic filtering can be added.

Bindings for *libmapper* exist for many popular programming languages and include C/C++, Java, Python, and Node.js. External objects for Max and Pure Data are also available.

3 Webmapper

Two other GUI applications, *Maxmapper* and *Vizmapper* existed prior to the development of *Webmapper*. The former was an interface implemented in the Max environment that provides display and manipulation of connections on the network based on a list representation, and served as the seminal example of a usable graphical tool to view and manipulate devices on the network. The latter was an exploration of alternative representations of larger and more complex

networks [8], and its existence also demonstrated that a different visual representation of the mapping can be running concurrently due to the distributed nature of the network. A basic command-line application was also made to manipulate and view mappings on the *libmapper* network[10].

Webmapper, as its name implies, is a browser-based application. Originally, the motivation to build a web-based application was to provide the ability to run the user interface within a browser on a variety of desktop and mobile platforms. Additionally, the frameworks and libraries for modern web development platforms support scalable development and deployment of visual user interfaces.

A highlight of *Webmapper* is that it provides more than a single view on the mapping structure, which allows multiple, as well as concurrent visual representations.

3.1 Architecture and Implementation

Webmapper is implemented using a Python back-end that serves two main functions. First, the server provides interfaces to the *libmapper* network and allows querying and modification of the state of running devices. Second, the server hosts the front-end HTML/JavaScript content and synchronizes the state of the network with the user interface. An overall architecture of *Webmapper* is shown in Figure 1.

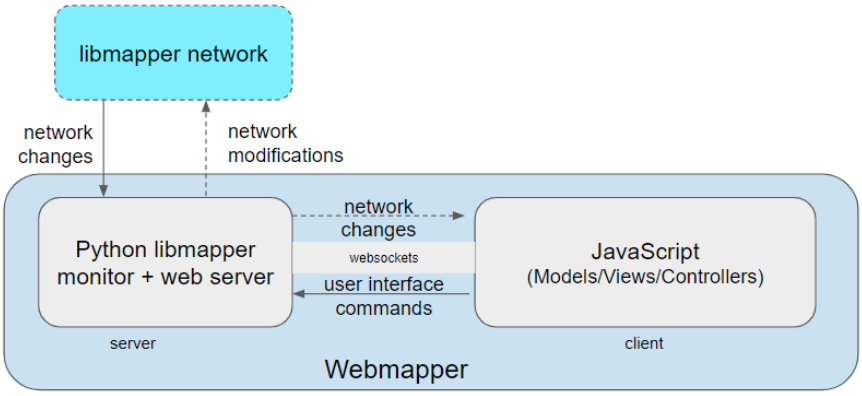


Fig. 1. Webmapper Architecture

[10] `https://sourceforge.net/projects/umapper/`

3.2 Views

The multiple views implemented in *Webmapper* were based on prior tools that had already been developed, as well as graphical design considerations pertaining to the correlation of properties of the network to various visual dimensions [9]. Each view provides a different method of visualizing and modifying the connections between devices and signals. Creation and modification of mappings are implemented via graphical input methods such as drag and drop between the visual elements, click to select, and keyboard shortcuts for removing connections. The full list of possible interactions for each view is described in [[9], chapter 4].

The following is a description of each view, followed by a visual demonstration example. The examples were created using saved mapping configuration files from a previous project, *Les Gestes: une nouvelle génération des instruments de musique numérique pour le contrôle de la synthèse et le traitement de la musique en performance par les musiciens et les danseurs*[11], a collaborative research project directed by Sean Ferguson and Marcelo Wanderley at McGill University and choreographer Isabelle van Grimde from the Montreal-based dance company Van Grimde Corps Secrets[12]. This project involving multiple wearable interfaces and a modular software synthesis system. At the time when these mappings were created, *Webmapper* had not yet been implemented so the only view available was the list based representation provided by *Maxmapper*. In this sample mapping there are two input devices connected to three output modules. The two input devices are identical wearable DMIs worn by dancers, and they control different parameters of three output devices simultaneously. One receiving device, a spatializer controller, is controlled by both input devices while two synthesizers are driven by each input device independently. Each view implemented allows a different way of visualising the connections in the network.

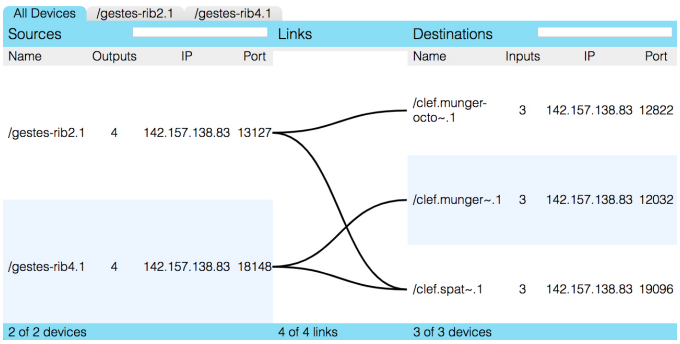


Fig. 2. List View device connections

[11] Gestures: a new generation of digital musical instruments for controlling synthesis and processing of live music by musicians and dancers

[12] `www.vangrimdecorpssecrets.com/`

List View The List View, one of the most direct ways of visually representing connections, simply provides a bipartite graph showing source devices on the left and destination devices on the right. Lines with arrowheads connect between source and destinations. Once a link is made through dragging and dropping between a source and destination device, a new tab window is created for the source device that allows signal to signal mappings to be made. Figure 2 shows the two input devices connected to 3 output devices. Here we can see that input device 1 is connected to output devices 1 and 3, while input device 2 is connected to output devices 2 and 3.

Selecting the tab window of input devices, we see how the individual signals are connected to the synthesizer inputs, as shown in Figure 3 left and right for the two devices.

The main advantage of the list view is that it lists all the connections between devices and signals at once. However, when there are a lot of connections, the visualization can become cluttered very quickly.

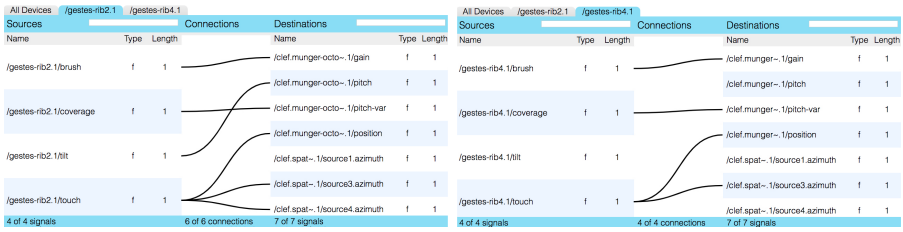


Fig. 3. List View signal connections for input device 1 and 2

Grid View In this view, inspired by the EagenMatrix[13] application, the network is represented by two grids. The left grid lists source devices on the horizontal axis and destination devices are on the vertical axis. Intersection points, if filled in blue, show the existence of links between devices. Devices must be added to the right grid to show their signals and connections; vertical/horizontal lines indicate the device has been added, and the right grid provides a similar representation for signals and connections. In Figure 4, only the first input and first output devices have been added. In figure 5, all devices have been added. The grid view is equipped with the ability to save view configurations into presets, allowing you to switch quickly between them.

One advantage of the Grid view is that, unlike the List View, a large number of connections will still be legible since there are no overlapping lines used to represent each link.

[13] `http://www.hakenaudio.com/Continuum/eaganmatrixoverv.html`

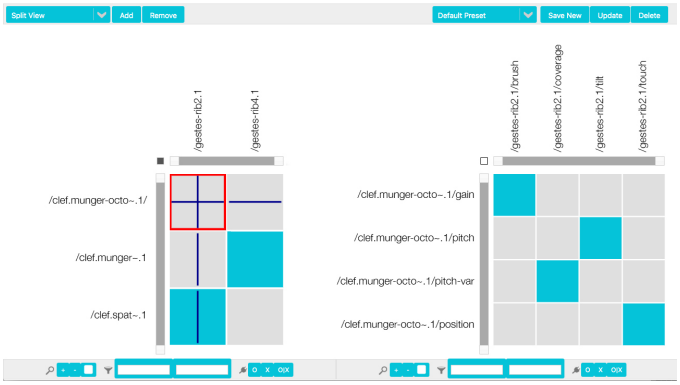


Fig. 4. The Grid View, showing connections between two devices

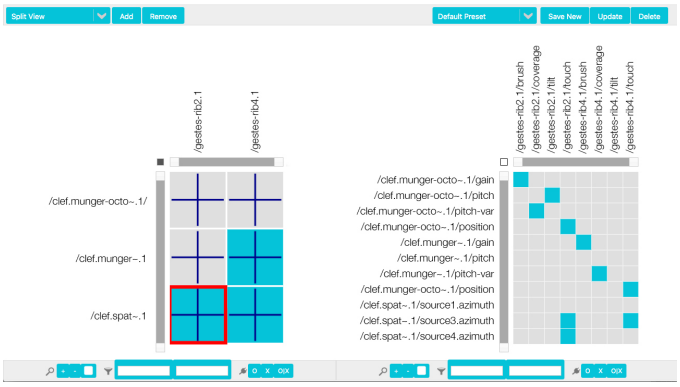


Fig. 5. The Grid View, showing connections between multiple devices

Balloon View Based on the tool implemented in [8], this view displays signals as a nested hierarchy—generated from textual analysis of the signals' OSC address URLs—of smaller circles within a larger one representing the device. Like the Grid View, it allows multiple source devices to be displayed at the same time (Figure 6).

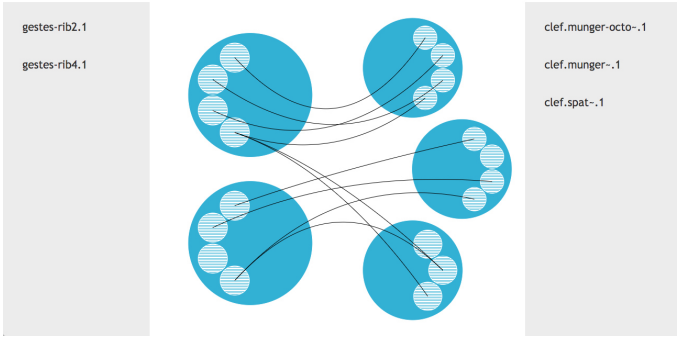


Fig. 6. The Balloon View

Unique to the Balloon View, a device can be selected by clicking inside the circle causing a "zoom" into the device, which then shows the individual signals as larger circles and lists the individual signals of the device on the side legend, providing further levels of detail. Figure 7 shows an input device 1 selected, followed by output device 1 in Figure 8. Clicking on the top of the device labels will "zoom out" of the selected device.

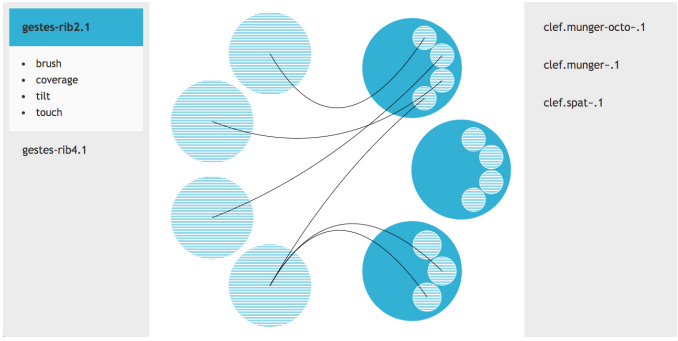


Fig. 7. Zooming in on a source device in the Balloon View

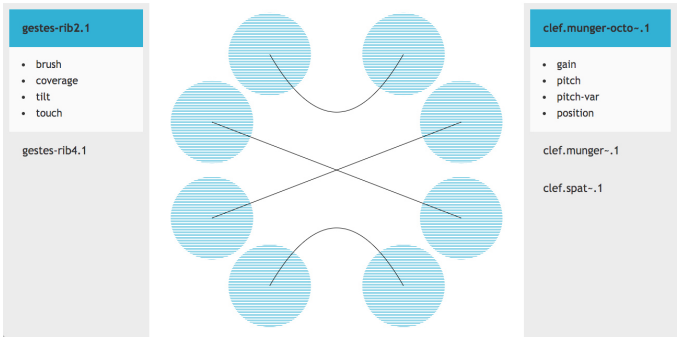


Fig. 8. Zooming in on both the source and destination

Hive View This view displays each device on a single axis, with nodes representing each signal. Lines between nodes on different devices show connections between signals. This view, like the Balloon View, allows multiple source devices to be displayed at the same time.

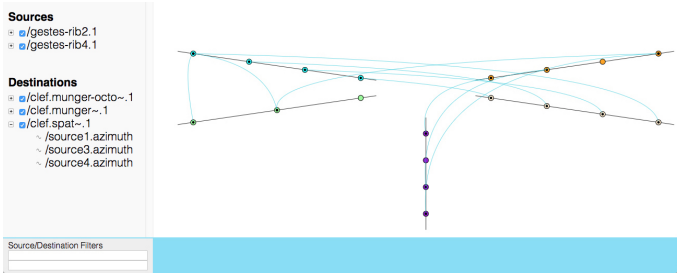


Fig. 9. The Hive View

Similar to the selective rendering features of other views, it is possible to filter out devices and signals. In the Hive View the selection is performed using checkboxes on the left. Signal names are displayed at the bottom of the screen when a particular connection is selected. Figure 10 shows these two features.

Connection Editor The above views provide different methods for visually representing and working with the basic signal connections. Since *libmapper* also provides processing built into the connection itself, a separate interface was implemented to display and modify the mapping expression and range parameters, as shown in Figure 11. This interface allows the signal processing features of *libmapper* to be viewed and modified. A basic interface for saving and restoring

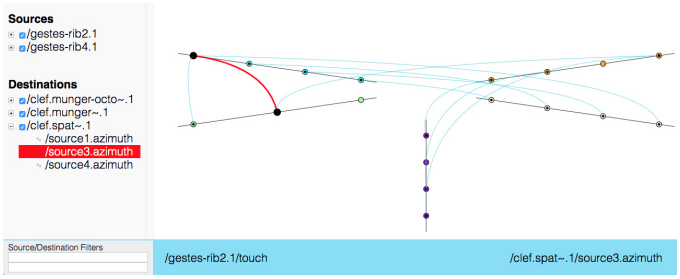


Fig. 10. Selective rendering and labeling in the Hive View

a mapping is also provided. The mapping configurations are saved as JSON[14] files, and when loaded, *Webmapper* will attempt to restore the recorded mapping connections onto the network by triggering the associated *libmapper* commands.

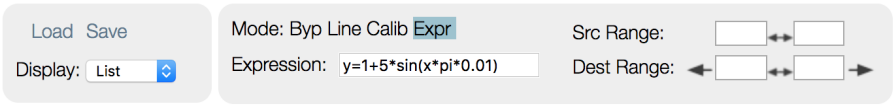


Fig. 11. The load/save and connection editor interface, showing an arithmetic expression applied to the connection.

3.3 User Evaluation

An informal user evaluation was performed on three different views (List, Grid, and Hive)[15] using measurable factors of *time to learn, speed of performance, error rate, and subjective satisfaction* [10]. Three users, who were researchers in the field of DMI design, participated in the evaluation. In the experiment the users were asked to create a mapping for DMIs that they were prototyping using the List View, erase it, and then recreate it using the Grid and then Hive views. The users were observed by the experimenter and followed up with a discussion to investigate the observations and collect general user feedback. A more comprehensive description and discussion on the evaluation is presented in Chapter 5 of [11]. Table 1 contains a summary of the rankings (1=best, 3=worst) of each view for the measured metrics.

The evaluation shows that the alternative interfaces provide quite different methods of interaction and visualization with strengths and weaknesses in different situations. For example, the Hive View was easiest to understand as entire

[14] JavaScript Object Notation
[15] The Balloon View was not yet implemented at the time of evaluation.

Table 1. Measurable human factors for each view

Metric	List	Grid	Hive
Time to learn	2	3	1
Speed	1	2	3
Error rate	1	2	3
Subjective Satisfaction	1	2	3
Ability to visualize	3	1	2

devices and signal connections were presented at the same time, but it was much harder to find a particular signal since it required selecting a signal node before the name of the signal can be revealed. The List View, on the other hand, provided the most straightforward visual representation of connections, but due to the separate tabs for each output device, did not provide for a easy way to obtain an overall picture of the network without switching tabs.

4 Discussion and Future Work

The results of the described evaluation provide a starting point in showing the differences between each view based on a limited number of metrics for a single user performing very specific tasks. In order to further justify the original motivations, additional in depth evaluations should be done to qualify the success in which these different representations have on the original motivations of supporting experimentation and diversity of users. Since the system provides multiple, concurrent representations of the network, the third goal of distributed control is therefore fulfilled by the nature of the implementation.

In terms of the visualizations themselves, thus far we have only implemented the representation of the mapping structure. However, the process of DMI design goes beyond just the connection between signals at a certain point in time: For example, how the mappings are modified over time, as well as the actual signals transmitted through the network are additional factors worth consideration. Integration of version control [12] into the tool as well as live signal visualization (in tools like OSCulator) and analysis can be useful for recalling the temporal progression of the design process, and afford further insight on the current state of the system, respectively.

5 Conclusion

In this paper we have presented *Webmapper*, a visual tool for viewing and manipulating mappings in the context of DMI design. By providing a visual interface for devices and signals on the *libmapper* network with multiple representations and interaction methods, we aim to support different DMI design concepts and workflows, especially in collaborative contexts. It should be stressed that although the views implemented in this application were motivated by specific

perspectives, the overall framework is intended to support the creation of tools to fit a diversity of perspectives. With completely open-source code and a multitude of language bindings it is relatively easy for developers to build additional tools for *libmapper*, and the modular nature of *Webmapper* supports the rapid addition of new views and manipulation strategies. We hope that further usage of the ecosystem demonstrated through the implementation of *libmapper* and *Webmapper* will lead to the development of more interesting perspectives on mapping, both through internal features as well as new tools and use-cases.

References

1. Miranda, E.R., Wanderley, M.M.: New digital musical instruments: control and interaction beyond the keyboard. Volume 21. AR Editions, Inc. (2006)
2. Hunt, A., Wanderley, M.M., Paradis, M.: The Importance of Parameter Mapping in Electronic Instrument Design. Journal of New Music Research **32**(4) (2003) 429–440
3. Steiner, H.C.: Towards a catalog and software library of mapping methods. In: Proceedings of the 2006 Conference on New Interfaces for Musical Expression. NIME '06, Paris, France, France, IRCAM - Centre Pompidou (2006) 106–109
4. Bevilacqua, F., Müller, R., Schnell, N.: MnM : a Max / MSP mapping toolbox. In: Proceedings of the International Conference on New Interfaces for Musical Expression. (2005)
5. Van Nort, D., Wanderley, M.M., Depalle, P.: Mapping Control Structures for Sound Synthesis: Functional and Topological Perspectives. Computer Music journal **38**(3) (2014) 6–22
6. Fiebrink, R., Trueman, D., Cook, P.: A metainstrument for interactive, on-the-fly machine learning. In: Proceedings of the International Conference on New Interfaces for Musical Expression. (2009) 280–285
7. Malloch, J., Sinclair, S., Wanderley, M.M.: Distributed tools for interactive design of heterogeneous signal networks. Multimedia Tools and Applications (2014) 1–25
8. Rudraraju, V.: A Tool for Configuring Mappings for Musical Systems using Wireless Sensor Networks. Masters thesis, McGill University (2011)
9. Krajeski, A.H.: A Flexible Tool for the Visualization and Manipulation of Musical Mapping Networks. Masters thesis, McGill University (2013)
10. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. 3rd edn. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1997)
11. Wilansky, J.: A Software Tool for Creating and Visualizing Mappings in Digital Musical Instruments. Masters thesis (2013)
12. Wang, J., Malloch, J., Chevalier, F., Wanderley, M.: Versioning and Annotation Support for Collaborative Mapping Design. In: Sound and Music Computing Conference. (2017)

The Tragedy Paradox in Music: Empathy and Catharsis as an Answer?

Catarina Viegas[1], António M. Duarte[2], and Helder Coelho[3]

[1] Colégio Mente-Cérebro, Universidade de Lisboa, Lisboa, Portugal
[2] Faculdade de Psicologia, Universidade de Lisboa, Lisboa, Portugal
[3] BioISI e Colégio Mente-Cérebro, Universidade de Lisboa, Lisboa, Portugal

Abstract. This paper suggests an hypothetical explanation for the Tragedy Paradox in music (i.e., the possible motivation to listening music that provokes negative feelings parallely to the tendency to avoid negative experiences). On the basis of a literature review on neurobiology, philosophy of mind, psychology of art and social cognition, it is proposed that listening to sad music might promote a rewarding cathartic process, eventually driven by empathy based on mirror-neurons system activity and cognitive empathy mechanisms.

Keywords: catharsis · cognitive empathy · emotional empathy · mirror-neurons system · tragedy

1 Introduction

The appealing nature of sad music has drawn the attention of philosophers through-out history from Aristotle [1] to Schopenhaur and it has been particularly empirically explored during the last decade [11], providing evidences in the fields of neuroscience and psychology [9]. The Tragedy Paradox in the context of music lays on the puzzling appreciation of music that provokes negative feelings when negative experiences tend to be avoided in everyday contexts [40]. Sadness is considered to be a typical short-term response to a personal loss, corresponding to a negative experience that, generally, people are lean to avoid [2, 32]. Sad music, on the other hand, is related with specific cues recognizable across cultures [27] and, although being considered able to activate the same machinery responsible for real-life emotions [9, 25, 30], the emotional experience that it induces varies within a spectrum of emotions that is intricate [32,41], such that the music that is reported has being able to induce sadness might be simultaneously described as pleasurable [9].

Philosophers have been suggesting that its appreciation when induced by music derives from the lack of real world consequences, a proposal that have inspired posterior explanations [35,38,41]. Huron [17], for instance, has offered an hypothesis specific to sadness based on the associated endocrine responses, which function as a mechanism of pain relief and are also susceptible to be triggered by sad music. Huron suggests that those responses, in the absence of a real loss when listening to music, lead to an overall positive experience. However, such

hypothesis has not been empirically verified yet. Moreover, listener's personality traits and the context where sad music is listened constitute relevant factors for the understanding musical preference. In that sense, Garrido and Schubert [13] propose that *empathy traits* predict a preference for sad music whereas other authors show that sad music tend to be listened when self-regulatory processes are demanded from the listener [42] or in solitary settings [41]. Importantly, higher levels of empathy traits are positively correlated with the cases where sad music is a source of consolation [9, 41, 42, 45] and is associated with a positive experience for the listener [22, 44].

Having into account how Aristotle [1] answered the same Paradox in the realm of Greek Tragedy, proposing *catharsis* as the process responsible for the tragic pleasures, once it guided the audience to purge inner negative feelings through its own experience, the present paper aims to present an hypothesis for the cathartic process that might takes place when listening to sad music driven by an empathetic engagement. In this regard, two questions arouse: first, how does empathy plays an active role for sadness to be felt with music and, second, how does empathy explain the rewarding side of such sadness? Although some previous mechanisms were already proposed by Levinson from a philosophical perspective [28], it has been recently shown that *cognitive empathy* is involved in musical processing [45], which represents an important evidence to defend *catharsis* as the main reward of sad music listening. This perspective on the Tragedy Paradox in music brings an original view to explain the rewardings associated with listening to sad music.

2 The Influence of Empathy in Musical Processing

It sounds puzzling at the first sight that someone is able to empathize with music, however, recent evidences [31, 45] show that musical processing and empathetic capacities rely on the shared neural circuits. The word empathy has origin in the Greek *empatheia*, translated by *em* - 'in' + *pathos* - 'feeling', meaning physical affection or passion, which was used to translate the German word *Einfühlung* that means "feeling into" [6]. The empathetic experience is distinguished according with *emotional* or *cognitive empathy* processes, whether they involve the involuntary capture of others' emotions or the conscious understanding of what others are thinking and feeling [6, 38]. The individual tendencies to behave according with these two types of processes reflect different personal traits measured according with the Interpersonal Reactivity Index (IRI) - *empathic concern* and *personal distress*; *fantasy* and *perspective taking* [45]. *Empathic concern* describes the automatic concern toward the others, whereas *personal distress* refers to the negative emotional response generated automatically when facing someone in a stressful situation - both representing *emotional empathy* traits. On the other hand, *perspective taking* represents the ability to shift one's perspective in order to understand others' situations, while *fantasy* corresponds to the tendency to imagine oneself in the position of others - being both representative of cognitive empathy [6]. From a neurological perspective, mirror-neurons system

have been discussed as a circuitry responsible for our ability to empathize, due to their discharge either when a person observes or performs an action, therefore leading to action understanding [31]. The discovery of mirror-neurons was first made in macaque brain [34], but such position is refuted by some authors [16] who argue that the activation of non-motor areas of the brain might be sufficient to comprehend a perceived action, defending that mirror system rather reflect sensory-motor associations. However, Kemmerer [24] shows how the simulation of actions by mirror-neuron system supports the understanding of actions through the decoding of "how" are they performed, representing a mechanism to access to the intentions of others [19, 31, 45] through *emotional empathy*. On the other hand, *cognitive empathy* might be responsible for the decoding of "why" are actions performed being particularly relevant when applied to social intentions, which is considered to be a mentalizing activity. Mentalizing is specifically associated with the activity of dorsomedial prefrontal cortex and the temporoparietal junction (TPJ) [24]. Mirroring properties are not only evidenced for movements but also for emotional states, which happens when facing others' pain situations by the consistent activation of anterior insula (AI), anterior cingulate cortex (ACC) and the inferior frontal cortex [3, 20, 26]. Mentalizing activities and mirroring properties of the brain constitute functions particularly connected, such that, mirror-neurons are candidates to simulate the mental states of others in order to understand what are they thinking or feeling [15, 23, 39].

From a psychological approach, *emotional empathy* is carried out through processes of *emotional contagion*, i.e. the internal simulation of emotions perceived [8], whether in facial expressions or gestures [21]. The internal simulation of an emotion represents its *internal locus*, whereas its external cause rather corresponds to its *external locus* [38], the shifting of the mental state from the internal to the external locus allows to manipulate bottom-up affective information attributing it to external sources through the integration of contextual information due to the intervention of top-down mechanisms. In turn, the attribution of mental states to external sources drives prosocial behavior [38], as it is clarified in Fig.1. Bottom-up mechanisms represent the activity of *emotional empathy* whereas top-down mechanisms rather reflect *cognitive empathy* processes. The greater tendency to activate mirror system, which is related with the proclivity to employ prosocial decision-making [5]. Even though *emotional* and *cognitive* empathy show a reasonable functional independence, both systems work interconnected. The involvement of mirror-neurons in both *cognitive* and *emotional empathy* [18, 39], supports the idea that low-level processes of simulation are at the base of higher-level processes [12] involved in executive control, regulation of emotions, mentalizing, contextual appraisal and enactment imagination [45]. What is the relation between these mechanisms and music listening?

2.1 How does the Sadness Felt when Listening to Music Depends on Empathy?

Indeed, Molnar-Szakacs and Overy propose that mirroring properties seem to be associated with auditory system [31], which are directly associated with *emo-*

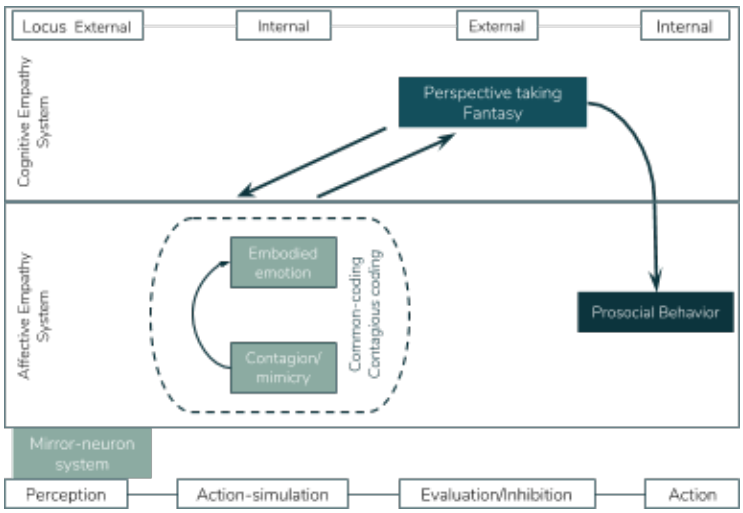


Fig. 1. Emotional (bottom-up) and cognitive empathy (top-down) mechanisms. Adapted from [38]

tional contagion processes through auditory stimuli positively correlated with a personal disposition for general empathy [41]. As Molnar-Szakacs and Overy highlight, the connection of mirror-neurons with the limbic system (involved in the processing of complex emotions), promote the internal simulation of emotions through the music [31], such that the sadness evoked by the music might be a reflection of the *emotional empathy* mechanisms. Mirror-neuron system constitutes a mechanism that responds automatically to musical signals due the similarity they have with voice-like aspects of emotional speech and to their [21], meaning that they constitute innate mechanisms independent from learning. Thus, the sadness expressed in the music emotion-specific auditory patterns activates its emotional representation in the brain, guiding sadness induction through *emotional contagion* [21]. On the other hand, *cognitive empathy* rather have an influence in the evocation of sadness through the processing of the lyrics' semantic content. Mechanisms of *imagination* are proposed to be involved in the conjuring of mental images associated with an emotional value responsible for activating the respective simulation Such mechanisms might be associated with the involvement of what António Damásio [7] called the *as-if-body-loops*, i.e. circuits in the brain responsible for the simulation of affective narratives experienced in the past that recreate feelings detached from its real experience or experiencing only a pale version of it.

As the Fig. 2 depicts, *emotional empathy* system drives the simulation of sadness through *emotional contagion* processes based on the activation of mirror-neurons according with the musical features, whereas *cognitive empathy* system rather drives processes of *imagination* to access the knowledge of what means to

be in a sad emotional state driven by the content of the lyrics, simulating states of *as-if-sadness*. Emotional process is considered to involve three main stages: The identification of a competent stimuli able to trigger the process; the generation of the emotional process according with its different sub-components (*subjective feelings, expressive medium, physiological arousal, action tendencies* [21]) and, lastly, the regulation of these responses in order to produce contextually appropriated behaviours [33]. These simulations of sadness constitute a way to localize internally (*internal locus*) the emotion caused by music through the activation of circuits involved in the generation of emotions. The *locus* is only transferred to the exterior by the employment of evaluations that constitute part of the regulatory behavior associated with an emotion. The result of such transference might be the appreciation of the object, in this case the music, resultant from emotions triggered from non-contagious mechanisms, again with an *intenal locus*. In the next sub-section it will be unveiled how these regulatory processes might be responsible for the rewarding of sad music listening.

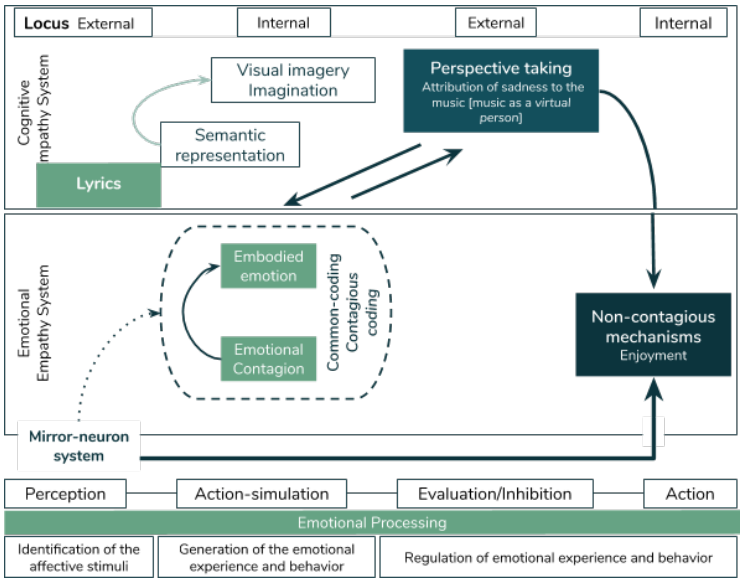


Fig. 2. Emotional and cognitive empathy mechanisms in music.

2.2 How Is Empathy Related with the Rewarding Side of Sad Music?

Some authors propose that the rewarding experiences associated with sad music derive from the lack of real consequences of sadness [11], once music is objectless

- it doesn't represent any loss for the self. However, the pleasurable side of negative emotions in music proposed here is beyond that lack of real consequences of sadness [9], being associated with a psychological benefit for the individual according with listeners' subjective reports [41]. How could sadness in music bring a pleasurable experience based on a psychological benefit?

The conclusions of a research carried by Wallmark and colleagues [45] suggest that sensor-motor engagement with music is not sensitive to the individual differences on empathy, meaning that emotional contagion is common among listeners. However, they show that the same does not happen relatively to the involvement of *cognitive empathy* areas in musical processing, since highly empathic people tended to show higher activation of areas involved in the processing of social information associated with *cognitive empathy*. The proposed hypothesis regards the role of *cognitive empathy* to overcome the emotional contagion effects and, indeedn listeners with higher levels of empathy tend to recruit areas of the brain involved in reward and motivation, such the limbic reward areas as the dorsal striatum and medial, lateral and orbital areas of the prefrontal cortex and TPJ when listening to music [45]. Moreover, musicians constitute a group of people that show increased emotional response to the musical signal - which turns them more susceptible to feel sad - but also an improved capacity to derive pleasure form sad music [4], revealing an higher activation of a network responsible for controlling emotional and motivational experiences - the striatal-thalamo-cortical loop [45]. This evidence suggests that they develop an enhanced affective neuroplasticity (homeostatic regulatory functions) as an adaptive counterpart of the repeated processing of negative emotions [4].

Together, these results suggest a cognitively mediated way of processing music [45] through which we are able to "feel into" as if it was another person, whether the composers' mind or the *virtual persona* of music. *Fantasy* traits activate different circuits according with the valence of the music (negative/ positive), generating pleasurable responses through their own identification with the musical emotions, whereas *perspective taking* traits tend to generate higher rewarding responses by the employment of modulatory mechanisms that minimize the negative experience of a contagious experience [21, 45]. The involvement of effortful, conscious and imaginative processes provide evidence for the way music is perceived: not only as a stimuli capable of automatic emotional responses, but also as a stimuli to which the listener is prone to attribute meaning. In this sense, music do not differ from any other social stimuli. The employment of such cognitive mechanisms of *empathy* constitutes the root for the shift from the internal to the external *locus* of sadness, which means that sadness is attributed to the music itself or to the people behind the music, rather than becoming property of the listener.

3 *Catharsis* in Music as a Function of Empathy

Considering the role of *cognitive empathy* mechanisms, they are very strong candidates for the transformation of negative into positive emotions and, therefore,

for the cathartic process that might take place when listening to sad music. Considering that the main reasons that take people to listen to sad music are emotional distress situations (failures, frustrations, breaking-ups or death) and moments of loneliness, sad music becomes a source of comfort and consolation through mood-sharing, i.e. by sharing the same emotional state with the listener and from there by establishing virtual social contact (music playing the role of a friend) [11, 13, 41, 42]. The consequences of employing circuits important for social processing associated with *cognitive empathy* allow to understand sad music as an anonymous subjectivity through which is possible to feel identified with. In this sense, establishing an empathetic relationship with the music becomes a rewarding activity for the listener.

According with Aristotle [2010, as cited in 19], a negative emotional response allows one to "bleed off in a controlled manner" purging a certain amount of negative emotions, which function as a momentarily painful dose of "emotional medicine". The cathartic process proposed for the specific case of sad music corresponds to the set of specific rewards that take place when listening to it. The first reward might result from the acknowledgement of the musical context where emotions take place, corresponding to the previously described *no real life implications rewards*, which derive from the *perspective taking* modulations. Afterwards, also through the ability to take perspective, it might become possible to recognize the emotional expressivity of the piece, which might be achieved by the change of locus of the emotion to the music and justifies the *apprehending emotions rewards* proposed by Levinson [28]. This reward stems from the cultural value of music, through which sadness becomes a valid emotion for the listener. The third phase of this process involves *emotional communion rewards* , which results from the comprehension of sadness as a shared emotion, whether by employing *perspective-taking* traits with the virtual persona of music or the composer or by the identification with the sadness expressed in the music through the employment of *fantasy* traits.

All these rewards end up in *catharsis*, constituting the building blocks of *cathartic rewards* through which the sadness felt with the music is transformed into a valuable psychological comfort, as Fig. 3 clarifies. Sad music, particularly, prompts higher activity in the right dorsolateral prefrontal cortex and in the right parietal areas and TPJ, both areas involved in regulatory processes and *cognitive empathy* functions [4,5]. Thus, the ability to regulate emotions becomes crucial for the overcoming of the automatic negative experiences felt when listening to music. The modulation of such emotional response by social cognition faculties helps to understand how feelings of *being moved* are important for the appreciation of sad music [44]. In turn, the rewarding effects of sad music are very much dependent on the personal contexts where it is heard, which become strong predictors for the motivation to recur o sad music in similar moments of personal distress or loneliness.

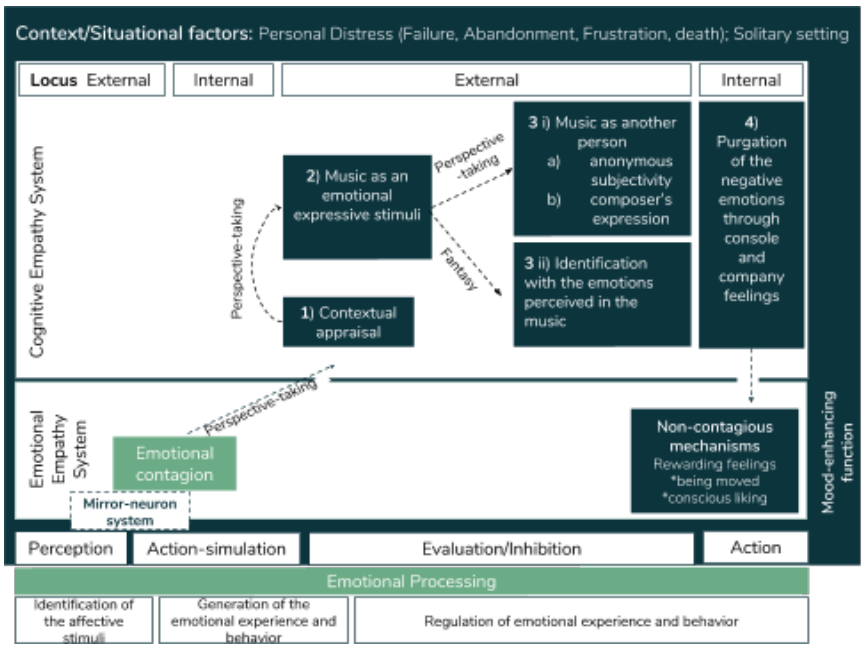


Fig. 3. The influence of cognitive empathy for the appreciation of sad music.

4 Conclusion

According with a literature review in neurobiology, philosophy of mind, psychology of art and social cognition, it is proposed an original answer for the Tragedy Paradox based on a cathartic process. This process is considered to be fully moderated by empathy: at a primary level as a mean to capture the emotional content of music, either driven by *emotional contagion* or *imagination* mechanisms; at a secondary level, as a mean to get rewards from the music guided by *cognitive empathy* mechanisms that allow the overcoming of the initial negative emotions. The regulation emotional states in contexts where the regulation of emotions is meaningful for the listener becomes, then, the main source of reward. Thus, music represents a non-invasive tool to stimulate the brain in areas involved in homeostatic regulation of emotions and *cognitive empathy*, which constitutes a topic to be further investigated in musical therapy contexts, where music might constitute a complementary tool in patients who show psychiatric disorders associated with losses or imbalances in social functions and in the regulation of emotional states.

5 Acknowledgements

We thank to David Yates for his supervision of the work.

References

1. Aristotle (2010). Poetica, (Eudoro de Sousa, trad.). Lisboa: Imprensa Nacional - Casa da Moeda.
2. Bonanno, G. A., Goorin, L., and Coifman, K. G. (2008). Sadness and Grief. In M. Lewis, J. Haviland-Jones, and L. Feldman Barrett (Eds.), The Handbook of Emotions (pp. 797–810). New York: Guilford.
3. Botvinick, M., Jha, A. P., Bylsma, L.M., Fabian, S.A., Solomon, P.E., Prkachin, K.M. (2005). "Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain". NeuroImage. 25 (1), 312–319. doi:10.1016/j.neuroimage.2004.11.043. PMID 15734365
4. Brattico, E., Bogert, B., Alluri, V., Tervaniemi, M., Eerola, T., and Jacobsen, T. (2016). It's Sad but I Like It: The Neural Dissociation Between Musical Emotions and Liking in Experts and Laypersons. Frontiers in Human Neuroscience, 9, 676. http://doi.org/10.3389/fnhum.2015.00676
5. Christov-Moore, L., Sugiyama, T., Grigaityte, K., and Iacoboni, M. (2017). Increasing generosity by disrupting prefrontal cortex. Soc. Neurosci. 12, 174–181. doi: 10.1080/17470919.2016.1154105
6. Clarke, E., DeNora, T., and Vuoskoski, J. (2015). Music, empathy and cultural understanding. Physics of Life Reviews, 15, 61-88. http://doi.org/10.1016/j.plrev.2015.09.001
7. Damasio, A. (2017). A estranha ordem das coisas. Lisboa, Círculo de Leitores.
8. De Waal, F. B. (2007) The 'Russian doll' model of empathy and imitation. In S. Bråten (Eds.) On Being Moved. From Mirror Neurons to Empathy (pp. 49-69). Amsterdam/Philadelphia: John Benjamins Publishing Company.
9. Eerola, T. and Peltola, H. R. (2016). Memorable experiences with sad music-reasons, reactions and mechanisms of three types of experiences. PLoS ONE, 11(6). http://doi.org/10.1371/journal.pone.0157444
10. Eerola, T., Vuoskoski, J. K. and Kautiainen, H. (2016). Being moved by unfamiliar sad music is associated with high empathy. Frontiers in Psychology, 7, 1176. http://doi.org/10.3389/fpsyg.2016.01176
11. Eerola, T., Vuoskoski, J. K., Peltola, H. R., Putkinen, V. and Schäfer, K. (2017). An integrative review of the enjoyment of sadness associated with music. Physics of Life Reviews, 25, 100-121. http://doi.org/10.1016/j.plrev.2017.11.016
12. Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and the neural basis of intersubjectivity. Psychopathology, 36(4), 171-80. http://doi.org/10.1159/000072786
13. Garrido, S., and Schubert, E. (2011). Individual Differences in the Enjoyment of Negative Emotion in Music: A Literature Review and Experiment. Music Perception: An Interdisciplinary Journal, 28(3), 279–296. http://doi.org/10.1525/mp.2011.28.3.279
14. Gazzola, V., Aziz-Zadeh, L., and Keysers, C. (2006). Empathy and the Somatotopic Auditory Mirror System in Humans. Current Biology, 16(18), 1824–1829. http://doi.org/10.1016/j.cub.2006.07.072
15. Goldman, A. I. (2013). Two Routes to Empathy: Insights from Cognitive Neuroscience. In Joint Ventures: Mindreading, Mirroring, and Embodied Cognition (pp. 198-217). Oxford, UK: Oxford University Press. http://doi.org/10.1093/acprof:osobl/9780199874187.003.0009
16. Hickok G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. Journal of cognitive neuroscience, 21(7), 1229–1243. doi:10.1162/jocn.2009.21189

17. Huron D. (2011). Why is sad music pleasurable? A possible role for prolactin. Music Science, 15(2), 146–58.
18. Iacoboni, M. (2012). The human mirror neuron system and its role in imitation and empathy. In F. B. M. de Waal and P. F. Ferrari (Eds.), The primate mind: Built to connect with other minds (pp. 32-47). Cambridge, MA, US: Harvard University Press.
19. Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., and Mazziotta, J. C. (2005). Grasping the intentions of others with one's own mirror neuron system. PLoS Biology 3(3), e79. http://doi.org/10.1371/journal.pbio.0030079
20. Jabbi, M., Swart, M., and Keysers, C. (2007). "Empathy for positive and negative emotions in the gustatory cortex". NeuroImage. 34 (4): 1744–53. doi:10.1016/j.neuroimage.2006.10.032.
21. Juslin, P. N., and Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. Behavioral and Brain Sciences, 31(5), 559-575. http://doi.org/10.1017/S0140525X08005293
22. Kawakami, A., and Katahira, K. (2015). Influence of trait empathy on the emotion evoked by sad music and on the preference for it. Frontiers in Psychology, 6(1541), 1541. http://doi.org/10.3389/fpsyg.2015.01541
23. Keysers, C. and Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. Trends in Cognitive Sciences, 11 (5), 194–196. doi:10.1016/j.tics.2007.02.002.
24. Kemmerer, David. (2015). Does the motor system contribute to the perception and understanding of actions? Reflections on Gregory Hickok's The myth of mirror neurons: The real neuroscience of communication and cognition. Language and Cognition. 2015. 450-475. 10.1017/langcog.2014.36.
25. Koelsch, S. (2014). Brain correlates of music-evoked emotions. Nature Reviews Neuroscience, 15(3), 170-180. http://doi.org/10.1038/nrn3666
26. Lamm, C., and Singer, T. (2010). The role of anterior insular cortex in social emotions. Brain Structure and Function, 214 (5-6), 579-91. http://doi.org/10.1007/s00429-010-0251-3
27. Laukka, P., Eerola, T., Thingujam, N. S., Yamasaki, T., and Beller, G. (2013). Universal and culture-specific factors in the recognition and performance of musical affect expressions. Emotion, 13(3), 434-449.
28. Levinson, J. (2011). Music, Art and Metaphysics - Essays in Philosophical Aesthetics, Oxford. UK: Oxford University Press.
29. Menninghaus, W., Wagner, V., Hanich, J., Wassiliwizky, E., Kuehnast, M., and Jacobsen, T. (2015). Towards a psychological construct of being moved. PLoS ONE, 10(6), e0128451. http://doi.org/10.1371/journal.pone.0128451
30. Mitterschiffthaler, M. T., Fu, C. H. Y., Dalton, J. A., Andrew, C. M., and Williams, S. C. R. (2007). A functional MRI study of happy and sad affective states induced by classical music. Human Brain Mapping, 28(11), 1150–1162. http://doi.org/10.1002/hbm.20337
31. Molnar-Szakacs, I., and Overy, K. (2006). Music and mirror neurons: from motion to "e"motion. Social Cognitive and Affective Neuroscience, 1(3), 235–241. http://doi.org/10.1093/scan/nsl029
32. Peltola, H. and Eerola, T. (2016) 'Fifty shades of blue : classification of music-evoked sadness.', Musicae scientiae., 20 (1). pp. 84-102.
33. Phillips, M. L. (2003). Understanding the neurobiology of emotion perception: Implications for psychiatry. British Journal of Psychiatry, 182(3), 190–192. http://doi.org/10.1192/bjp.182.3.190

34. Rizzolatti G, Fogassi L, and Gallese V. (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. Nat Rev Neurosci.; 2(9), 661–670.
35. Sachs, M. E., Damasio, A., and Habibi, A. (2015). The pleasures of sad music: a systematic review. Frontiers in Human Neuroscience, 9(9), 404. http://doi.org/10.3389/fnhum.2015.00404
36. Salimpoor, V. N., Benovoy, M., Longo, G., Cooperstock, J. R., and Zatorre, R. J. (2009). The rewarding aspects of music listening are related to degree of emotional arousal. PLoS ONE, 4(10), e7487. http://doi.org/10.1371/journal.pone.0007487
37. Schubert E. (1996) Enjoyment of negative emotions in music: an associative network explanation. Psychology Music, 24(1), 18–28.
38. Schubert, E. (2017). Musical Identity and Individual Differences in Empathy. In R. MacDonald, D. J. Hargreaves and D. Miell (Eds.) Handbook of Musical Identities (pp. 322-342). Oxford, UK: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199679485.003.0018
39. Shamay-Tsoory, S. G. (2011). The neural bases for empathy. Neuroscientist, 17(1), 18–24. http://doi.org/10.1177/1073858410379268
40. Smuts, A. (2009). Art and Negative Affect. Philosophy Compass, 4(1), 39–55. http://doi.org/10.1111/j.1747-9991.2008.00199.x
41. Taruffi, L., and Koelsch, S. (2014). The paradox of music-evoked sadness: An online survey. PLoS ONE, 9(10), e110490. http://doi.org/10.1371/journal.pone.0110490
42. Van Den Tol, A. J. M., and Edwards, J. (2015). Listening to sad music in adverse situations: How music selection strategies relate to self-regulatory goals, listening effects, and mood enhancement. Psychology of Music, 43(4), 473–494. http://doi.org/10.1177/0305735613517410
43. Vuoskoski, J. K., and Eerola, T. (2012). Can sad music really make you sad? Indirect measures of affective states induced by music and autobiographical memories. Psychology of Aesthetics, Creativity and the Arts, 6(3), 204–213. http://doi.org/10.1037/a0026937
44. Vuoskoski, J. K. and Eerola (2017) The pleasure evoked by sad music is mediated by feelings of being moved. Frontiers in Psychology, 8, 439.
45. Wallmark, Z., Deblieck, C., and Iacoboni, M. (2018). Neurophysiological Effects of Trait Empathy in Music Listening. Frontiers in Behavioral Neuroscience, 12(66). http://doi.org/10.3389/fnbeh.2018.00066
46. Wallmark, Z., Iacoboni, M., Deblieck, C., and Kendall, R. A. (2018). Embodied listening and timbre: perceptual, acoustical and neural correlates. Music Percept. 35, 332–363. doi: 10.1525/mp.2018.35.3.332

'Visual music'? The Deaf experience
Vusicality and sign-singing

Sylvain Brétéché

PRISM (UMR 7061 - Aix-Marseille University/CNRS)

breteche@prism.cnrs.fr

Abstract. This article aims to consider the visual dimensions of music, based on the Deaf practices represented by the vusic and the sign-singing (song in Sign Language), seeking to think how they can bring to a de-normalized consideration of music, namely the *vusicality*.

Keywords: Deaf people; Deaf musical experience; Vusic; vusicality; Sign-singing.

1 Introduction

Culturally, Deaf people[1] define themselves as 'visual beings' and the specificities of their condition necessarily imply that their capacities for perceiving the reality rest on particularly on its visual and dynamic aspects. But beyond to specify only a characteristic feature of world perceptions and representations for the Deaf, the visible presents itself for the 'People of the Eye' - typical Deaf expression - as the founding principle for the development of artistic practices and, in this way, as the primary sense of all aesthetic experiences.

Rather than simply content with the visual or visible arts, Deaf also seize, in their own way, practices that may initially seem inaccessible to them, unreachable or even 'forbidden' in some cases, such as dance and more specifically music. With their cultural affiliation to a community rich to its specificities, Deaf produce an extra-ordinary music that goes beyond the ordinary conceptions of current musical practices, developing what they call the *vusic* – a contraction of visual and music. A music of the eye, for the eye, which gives to see abandoning the aural dimensions commonly established to define the musical experience. More cultural still, they

[1] In this article, we use the designation 'Deaf' with capital D which, as specified by Charles Gaucher "announces a quest for identity which falls into a very precise historicity and is stated in terms which seek to turn the deaf difference into a cultural particularity detached from the physical incapacity which stigmatizes it" [1, p. 17].
For information, according to the SIL International census and estimates (2019), there are 144 Sign Languages around the world. However, the number of native speakers of these Sign Languages remains difficult to establish formally but can be estimated around 10 million (information available via www.ethnologue.com).

develop a typically deaf practice of the song in sign language, the sign-singing, where the signifying gesture takes musical values, the words becoming a visual melody, silent and embodied.

This article aims to consider the visual dimensions of music, based on the Deaf practices represented by the vusic and the sign-singing, seeking to think how they can bring to a denormalized consideration of music, namely the *vusicality*.

2 "Attitudinal deafness"[2]: the Deaf visual specificity

In view of the sensory specificities that characterize them, the Deaf develop a specific relationship to the world, putting aside the auditory realities and focusing primarily on visual and bodily qualities; because as Oliver Sacks specified, the Deaf community is "a community adapted to another sensory mode" [3, p. 251]. It is recognized today that the absence or the deterioration of a sensory modality can lead to the development of other sensory modalities, and recent studies emphasize this Deaf visual specificity [4; 5; 6; 7].

Therefore, and in the words of Owen Wrigley, "deafness is primarily a visual experience" [8, p. 29], and the Deaf willingly take possession of this 'visible' specificity that represents their singular relationship to the world. As writes Yves Delaporte:

> Deaf culture is a visual culture. Because hearing people also have a sense of sight, it is not sure that there is not much in common in the use that each makes of their eyes. Their eye gaze is invested with language functions [...]. [9, p. 36]

The Deaf visual qualities are characteristic of the Deaf identity, because "if for hearing people, being Deaf is defined by not hearing, for the deaf, being deaf is defined by the fact that to be visual" [10, p. 29]. The eyesight is thus essential for the Deaf sensory modality of the world apprehension. In a paper devoted to the issue of the 'Deaf eye gaze' [11], Yves Delaporte is interested in this self-designation of the Deaf as 'being-visual' and states:

> There is a specifically deaf way to permanently immerse yourself in all that the world can bring as visual information. The eye gaze is never passive or at rest, it is constantly attracted by everything in motion [...].
> This extreme sensitivity to everything within the visual field reflects recurring behaviors in time and space that we must consider them for what they are: cultural characteristics. [11, p. 50]

For the Deaf, the visual plays a fundamental role in their experiences of the world, exceeding the simple function of sensitive expression becomes the main modality of understanding and realization of the real. In addition, the specificities of gestural languages, fundamentally embedded into a visual expression, emphazise the importance given to the visual field by the Deaf. Thus, in the words of Yves

[2] "The most basic factor determining who is a member of the deaf community seems to be what is called 'attitudinal deafness'. This occurs when a person identifies him/herself as a member of the deaf community, and other members accept that person as part of the community" [2, p. 4].

Delaporte, we approach “what it is for the deaf to be deaf: it is to have capacities that hearing people do not have” [9, p. 38]. Indeed, for the Deaf, their condition is not defined primarily from their 'losses' but their abilities. They do not primarily think itself like beings whose the auditory system is impaired, but rather as individuals whose visual system is particularly operative: “We are visuals: this is the self-definition of the deaf” [9, p. 50]. This first cultural representation leads to consider the 'Deaf world' as a visuo-centered universe opposing the audiocentrism characteristic of the hearing world.

Moreover, the Deaf are fundamentally ‘speech beings’. This is the main cultural feature of the Deaf identity, and the Sign Language formalizes the essential criterion of membership of the Deaf community. The latter is defined as a linguistic and cultural minority; Sign Language is the natural language of the Deaf, their language which “reflects the culture, the traditions and the way in which the individuals who use it to communicate see the world” [12, p. 61]. More than a mere means of communication, the Sign Language represents for the Deaf the physical and ideological support of their identity representations. It is from their linguistic specificity that the Deaf have affirmed throughout History their identity and that they have elaborated their community gathering. Thus, the Deaf identity develops around another norm, visuo-centered and deeply embodied, which defines their relationship to the real, but also to the music.

3 ‘*Vusicality*’: seeing music

Indeed, in the Deaf musical experiences, the visual occupies a fundamental place. The Deaf specificities making the sight the dominant reception to perceive the material realities, in the musical experience, the eye complements the impaired ear to give meaning to sound phenomena. As Claire Paolacci points out, the Deaf “have a highly developed visual listening” [13, p. 55]; in this way, the music agrees with another sensitive dimension and takes on a specific value, singularly expressed in the ‘musical’ paintings by the deaf painter Chuck Baird which illustrate this *music for the eyes*.

However, the sounds are not materially seen and remain elements to hear and to feel; in the Deaf consideration of musical reality, certain elements involved in the creation of sounds become carriers of musical qualities. The deep sensitivity to vibrations that animate the body of the Deaf [14; 15; 16] agrees also with the elements perceived by the eye, attentive to visible movements that animate - in music, for music or by music - the visual space. As Emmanuelle Laborit explains:

> The concert show influences me too. The effects of light, the atmosphere, the many people in the concert hall, they are also vibrations. I am conscious that we are all together for the same thing. The saxophone shining with golden flashes, it is fantastic. The trumpeters who inflate their cheeks. [17, p. 30]

Thus, the music exceeds its only sound dimension, the musician bodies and musical objects are invested with a profound significance for the realization of the musical experience. Separated from its ordinary nature, music is no longer simply an

Art that *is listened* but is primarily an Art that *is looked*. As an artistic activity, music is a living Art that is performed in live and the concert represents a fundamental dimension of musical reality both to the Deaf and to the hearing people. To attend a concert is to see music being performed and the visual dimension, which also concerns the hearing audience, assumes a deep musical signification for the Deaf audience. In the words of Pierre Schmitt, "When music becomes a show, it is also through an increased focus on the visual aspects of the live performance that the musical experience takes on a particular significance for the deaf" [18, p. 228].

Deaf musical listening is not only perception and feeling of sounds, but it is also and fundamentally visualization of dynamics and movements that participate in the creation of the sensitive environment. Thereby, in the Deaf musical experience "the sight is a sense that draws the sound" [19] and brings to sound reality a more concrete existence revealing another form of materiality. Because the eye is sensitive to movements and visual rhythms, and as the deaf musician Maati Hel Hachimi points out,

> the deaf are able to understand the rhythm, to feel it without hearing, if only visually. For example, the train that passes with the wheels turning, the subway windows that scroll: we know very well if it goes more or less quickly and we feel the rhythm of what we see. [13, p. 49]

Thus, the movement and rhythm of the visual elements contribute to animate the Deaf musical experience. In this sense, the movements of the musicians seem essential, both for their participation in the reception of sound (felt and seen) and for their fundamental involvement in the musical practice. The gesture produced by the instrumentalist participates in this way to realize the perception of the sound elements, by bringing them a concrete origin and by giving to the vibratory feeling a visual base. As Maïté Le Moël points out: "every gesture is the cause of a bodily perception of the sound vibrations transmitted by the musical instrument" [20, p. 52]. Thus, the gestuality gives meaning to the Deaf musical listening but also contributes to the understanding of the dimensions and qualities peculiar to the musical practices.

Indeed, like for the hearing people, the musical practice for the Deaf requires a perfect command of specific technical gestures for producing the musical sounds with the instrument; however, in the Deaf practices, the gesture also presents itself as the fundamental understanding support of the musical elements, by participating to determine the sound differences and to define the notions of nuance, intensity or even rhythm. In fact, "it is by the meticulous control of the gesture and by the fine analysis of the bodily perceptions [that the Deaf] can discover the different variations of the sounds [...] and apprehend the notions of intensity, duration, and of height" [20, p. 53], making musician gesture an essential element for the musical practice. For the perfect command of the gestuality leads in a first instance to a control of the body in the musical activity, but it also leads to an understanding of the bodily perception capacities of sounds. The gesture presents itself as "a preferred means to feeling sensations and integrating certain sensory data transmitted by a sound emission" [13, p. 36]. Therefore, the musical gesture makes it possible to realize the sound event on the basis of visual and corporal elements.

But the gestuality is also for the Deaf the basis of their communication modality, and the Sign Language participate to define a singular facet of the Deaf musical reality revealing a specific practice, the sign-singing.

4 The sign-singing: 'the body sings silently'

Real musical practice from the Deaf world, the sign-singing proposes a soundless expression of a verbal text in the form of a signed song, where the body carries the melodic and rhythmic values by the exploitation of a "choreographed Sign Language, abstract and poetic" [18, p. 222]. Beyond presenting a simple translation of a vocal song into Sign Language, the sign-singing is deeply invested with musical dimensions that transform the common practice of Sign Language. Here, the musical experience accords with the Deaf specificities: the melodicity takes the body as the production space of the musical expression, whereas the rhythmicity of the gesture exploits the visual space as the realization place of the musical event. The signed song performances synthesize the specificities of the Deaf musical reality: the visual modality and the embodied practice of the musical experience. Affirming part of their musical identity with this singular practice, the Deaf distort the ordinary codes of the singing to produce a visual music which borrows the expressive values of the vocal to develop an exclusively bodily song. The sign-singing is, in a way, a silent musical expression, the silence of the Deaf expressed through the body like musical expressiveness support.

4.1 Musical parameters of sign-singing

The musical qualities of the sign-singing are close to ordinary musical parameters, although using them in specific ways according to the Sign Language characteristics. In this way, we can identify 5 criteria [21] that allow us to consider the musical dimensions of a signed song performance:

4.1.1 *The rhythmicity of language*

We find in the sign-singing a rhythmic transformation of the signs production; in a musical situation, these are indeed produced with a particular movement, which exploited the discourse energy with a specific dynamic more structured and orderly but less natural than the spoken communication.

The rhythm is also integrated into the whole body, which characterizes the global musical dynamics and animates the gestural production of lyrics. During the gestural communication, the body is not engaged in regular movements and is often adapted to the gestural specificities to the signs produce. In a musical situation, it is more specifically invested with marked movements that bring to the gestural production a particular expressiveness and give it these aesthetic qualities.

4.1.2 *Embodied melodicity*

The melodicity of a signed song rests on the development of the gestures in the communicational space and on the enlargement of signs production framework. In everyday gestural communication, Sign Language involves mostly the upper body - above the waist - in defined proportions; the 'sign space' designates the signs production sphere in the spoken communication, which defines

> The space surrounding the signer and that is reachable by these two hands. The sign space is used to locate the entities or notions associated with certain signs, possibly to specify their shape and size properties and to establish the spatial relations between the entities. [22, p. 220]

The sign space thus reports a specific area on the front of the signer's body, mainly between the shoulders and the waist. Forming "roughly a volume with a depth, a width and a height equal to the length of the speaker's arms" [23, p. 9], it defines the communicational framework of signs realization.

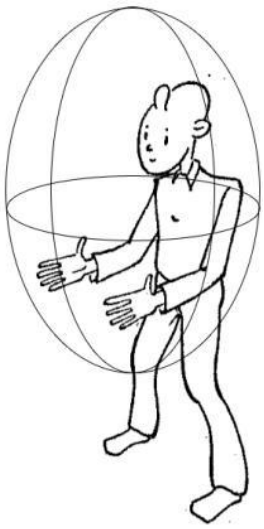


Figure 1. « sign space » [24]

The sign-singing, in its musical exploitation of Sign Language parameters, broadens the communicational sign space proposing an enlargement in height, width, and depth of signs production. The amplitude of signed song performances thus distinguishes the spoken production from its musical expression, bringing to the discourse its melodic form. The melody of sign-singing stands out from ordinary conceptions of the melodicity, which associate it with a succession of notes and pitch producing a characteristic and identifiable sound movement. In a signed song, the melody is coming from a movement, not a sounding movement but a visual

expression; the dynamics succession of signs produces a silent melody based on a specific usage of the sign place in a poetic way.

4.1.3 *Nuances and intensities*

The sign-singing is based on nuances, which do not appear here as sound qualities but as dynamic intensities. Rhythmicity and melodicity of the gestures are associated with a diminution or an enlargement of the verbal signs, formal transformations that intensify the musicality of the performance defining its aesthetic qualities. In a musical context, the body extends or reduced giving to the signs significant values, a phenomenon that is also found, to a lesser extent, in current gestural communication. Indeed, the sign-singing intensifies the expressive dimensions inherent in Sign Language, in order to requalify them into musical elements.

4.1.4 *Nuances and intensities*

The repetition process is significantly used in the sign-singing, firstly to add an expressive effect, but also to inject dynamism into the musical performance or accentuate its rhythmicity. It is common to find repeated signs, sometimes several times in a row, in a purely visual aesthetic perspective that transforms the gestural expression into a musical interpretation.

4.1.5 *Transposition of signs*

Finally, we find a transposition of the usual form of the verbal signs, which can sometimes be modified in their production (gestures enlargement or reduction; the speed of execution; production delocalized in the sign space) or totally transformed to perform the lyrics in a visual or poetic way (close to mime).

For example, in the Signmark's song *Against the Wall* [25], performed in American Sign Language (ASL) by the Finnish sign-singer, we can to find a formal transposition of the sign [WALL], whose usual configuration in ASL [Figure 2a] is transformed in a mimetic expression in the sign-singing execution [Figure 2b].

We can see that the musical using of the verbal sign [WALL] (hands side by side on the front of the body, which separate laterally at the shoulder width) transforms its initial disposition (in Signmark's performance, the hands are not side by side on the front of the body in the center of the torso, but at the shoulders close to the body) and its final resolution is extended (the hands do not stop at shoulders; the arms are outstretched). This transposition of the sign agrees with the expressive orientation of lyrics: "against the wall", words that the sign-singer performs physically.

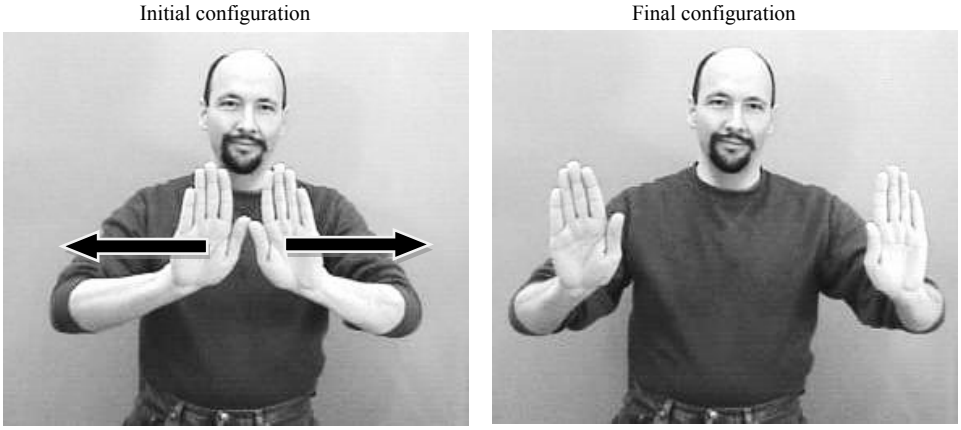


Figure 2a. [WALL] in ASL [26]

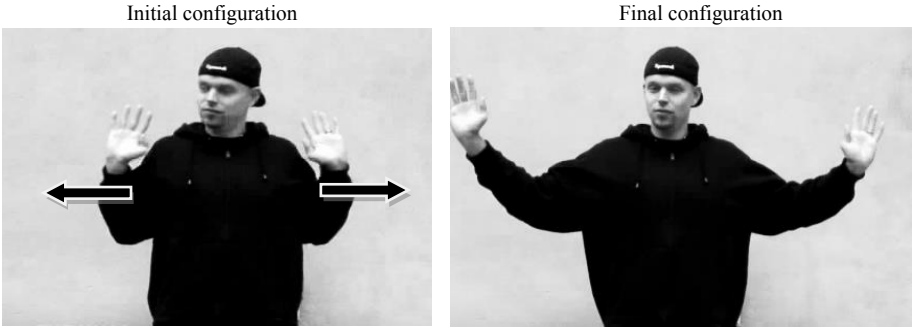


Figure 2b. [WALL] in Signmark's performance [27]

We can take another example of musical transposition, more explicit this time. In the same Signmark's song, the production of the verbal sign [WORLD] is totally detached from the usual sign [Figure 3a.] to be closer to a formal expression of the World [Figure 3b], formal expression that the linguistics of Sign Language calls the 'highly iconicity' [28], namely the insertion into the language of "structural indications of an illustrative representation of the sensory experience" [29, p. 23].

In his signed song, Signmark uses two configurations to perform [WORLD], which stand out from the usual sign. The first [Figure 3c.] is related to the round shape of the World and in no way to the verbal sign [WORLD]. In ASL, the sign [WORLD] is based on a configuration reflecting a low degree of iconicity (few illustrative values) and consists of two 'W' (the form of the hands into the manual alphabet), which rotate around each other to symbolize the Earth's rotation. In this, this sign has little iconic dimensions but refers more specifically to the word itself in its writing.

Figure 3a. [WORLD] in ASL [26]

Figure 3b. [WORLD]: formal representation

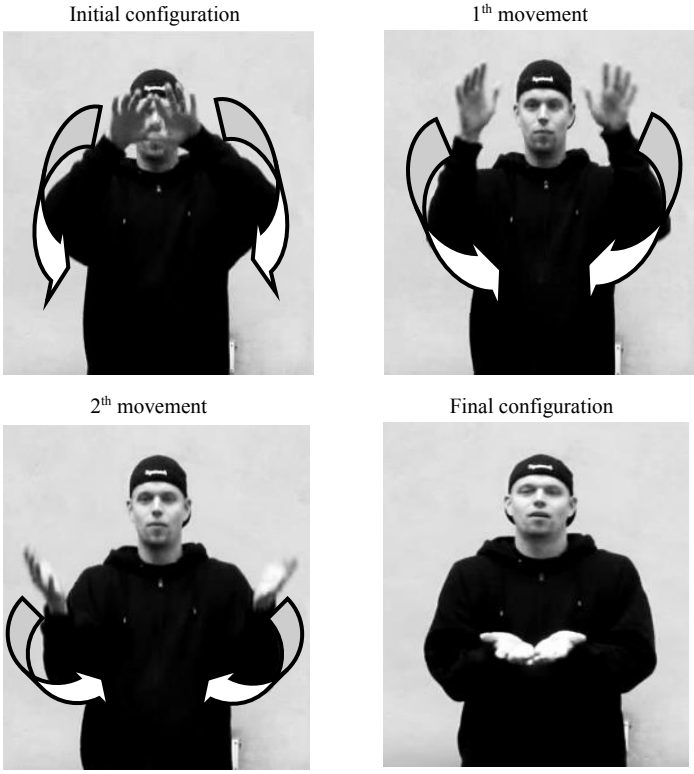
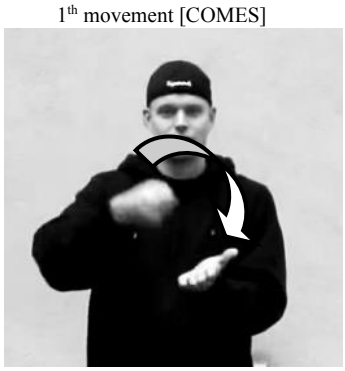


Figure 3c. [WORLD]. First expression by Signmark [27]

In the second evocation of [WORLD], the round shape evoking the Earth is transformed into a new expression [Figure 3d.], which presents again the shape of the world without referring to the usual verbal sign. This second expression is again part of an expressive process; in order to musically produce the lyrics "Even if the world comes crashing down", Signmark formalizes and choreographs the Sign Language to make visible his poetic and musical intentions. The WORLD is represented in its round shape (with the clenched fist) and this expression is used to express the lyrics in a mimetic way: the clenched fist 'crashes on' the hand. The expressiveness of the lyrics leads to transform the usual dimensions of Sign Language to bring a concrete and illustrative dimension to the sign-singing performance.

Initial configuration [WORLD]

1th movement [COMES]



Final configuration [CRASHING DOWN]

Figure 3d. [WORLD COMES CRASHING DOWN]
Second expression by Signmark [27]

*

The sign-singing thus reveals the silent appropriation of the musical codes of the ordinary song, adapting its expressive modalities to the Sign Language specificities. The music then becomes specifically Deaf and reveals the culture that defines the Deaf community, offering a singular way to claim a 'musical otherness'. The sign-singing involves the Sign Language in a musician practice that, transcending the ordinary norms of the song, revalorizes the notion of silence: *by the hands, for the eyes*, the sign-singing becomes a visual expression of music. So, we can consider with Pierre Schmitt that

> The musical experience claimed by the deaf exceeds the only sound sphere. Its meeting with the Sign Language poses the visual dimension not as a further or an additional dimension, but as a constitutive value of a musical form whose conception is enlarged. [18, p. 229]

More than just a communication mode, the Sign Language unveils aesthetic qualities that lead to the realization of original Deaf music, revealing singular creative perspectives and a strong musical identity, embodied and integrating primarily visual dimensions. Thus, the Deaf practices, by revealing the 'vusical qualities' of the music, make it possible to relocate the current conceptions of the music and offer to think the musical in its multimodal dispositions: the ear, but also the eye and more broadly the body proposing to concretize, together and jointly, the musical experience.

References

1. Gaucher, C.; Les Sourds : aux origines d'une identité plurielle. Pieter Lang, Bruxelles (2010)
2. Baker, C.; Padden, C.: American Sign Language: a look at its history, structure and community. T.J. Publishers, Silver Spring Md (1978)
3. Sacks, O.: Des yeux pour entendre. Voyage au pays des sourds. Seuil, Paris (1990)
4. Codina, C.; Buckley, D.; Port, M.; Pascalis, O.: Deaf and hearing children: a comparison of peripheral vision development. In: Developmental Science, Vol. 14, 4, (2011)
5. Stivalet, P.; Moreno, Y.; Richard, J.; Barraud, P.-A.; Raphel, C.: Differences in visual search tasks between congenitally deaf and normally hearing adults. In: Cogn. Brain Res., 6, pp. 227-232 (1998)
6. Parasnis, I.; Samar, V.J.: Parafoveal attention in congenitally deaf and hearing young adults. Brain Cogn. 1985. 4, 313–327.
7. Proksch, J.; Bavelier, D.: Changes in the Spatial Distribution of Visual Attention after Early Deafness. In: J. Cogn. Neurosci., 14, pp. 687-701 (2006)
8. Wrigley, Owen, the politics of deafness, Gallaudet University Press, Washington (1996)
9. Delaporte, Y.: Les sourds, c'est comme ça : ethnologie de la surdimutité. Éditions de la Maison des sciences de l'homme, Paris (2002)
10. Lachance, N.: Territoire, transmission et culture Sourde. Perspectives historiques et réalités contemporaines. Presses de l'Université de Laval, Québec (2007)
11. Delaporte, Y.: Le regard sourd. « Comme un fil tendu entre deux visages…". In: Terrain, n°30, pp. 49-66 (1998)
12. Dubuisson, C.: Signer ou le sort d'une culture. In: Nouvelles pratiques sociales, vol. 6, n°1, pp. 57-68 (1993)

13. Cité de la musique (ed.): Journée d'étude professionnelle Musique et surdité. Cité de la Musique, Paris (2005) www.citedelamusique.fr/pdf/handicap/260305_musique-et-surdite.pdf
14. Cranney, J.; Ashton, R.: Tactile spatial ability: Lateralized performance of deaf and hearing age groups. In: J. Exp. Child Psychol, 34, pp. 123-134 (1982)
15. Levänen, S.; Jousmäki, V.; Hari, R.: Vibration-induced auditory-cortex activation in a congenitally deaf. In: adult. Curr. Biol., 8, pp. 869-872 (1998)
16. Ammirante, P.; Russo, F.A.; Good, A.; Fels, D.I.: Feeling Voices. In: PLoS One, 8 (2013)
17. Laborit, E.: Le cri de la mouette. Robert Laffont, Paris (1994)
18. Schmitt, P.: De la musique et des sourds. Approche ethnographique du rapport à la musique de jeunes sourds européens. In: Bachir-Loopuyt, T., Iglesias, S., Langenbruch, A., & Zur Nieden, G. (eds.), Musik – Kontext – Wissenschaft. Interdisziplinäre Forschung zu Musik / Musiques – contextes – savoirs. Perspectives interdisciplinaires sur la musique. pp. 221-233. Peter Lang, Frankfurt am Main (2012)
19. Boyer, M.: La musique chez les enfants sourds. Mémoire du CAAPSAIS Option A (2001) http://atelieroptiona.free.fr/accat/accate.htm
20. Le Moël, M.: L'univers musical de l'enfant sourd. In: Marsyas n°39/40, Dossiers Pédagogies et Handicaps. pp. 51-58 (1996)
21. Brétéché, S.: L'incarnation musicale. L'expérience musicale sourde. Thèse de doctorat en musicologie. Esclapez, C., et Vion-Dury, J., (dir.). Aix-Marseille Université (2015)
22. Ben Mlouka, M.: Analyse automatique de discours en langue des signes : Représentation et traitement de l'espace de signation. In: Actes de la conférence conjointe Jep-Taln-Recital, vol. 3, Grenoble, pp. 219-232 (2012)
23. Segouat, J.: Modélisation de la coarticulation en Langue des Signes Française pour la diffusion automatique d'informations en gare ferroviaire à l'aide d'un signeur virtuel. Thèse de doctorat en informatique. Braffort, A. (dir.). Université Paris XI (2010)
24. Guitteny, P.; Legouis, P.; Verlaine, L.: La langue des signes. Centre d'Information sur la Surdité d'Aquitaine (2004)
25. Signmark, *Breaking the rules*, Warner Music (2010)
26. American Sign Language University (ASLU) http://www.lifeprint.com/asl101/
27. https://www.youtube.com/watch?v=JYOYvjhzy84
28. Sallandre, M-A.; Cuxac, C.: Iconicity in Sign Language: A Theoretical and Methodological Point of View. In: Wachsmuth, I.; Sowa, T. (ed.): Gesture and Sign Languages in Human-Computer Interaction. Springer, Berlin- Heidelberg-New-York pp.173-180 (2001)
29. Sallandre, M-A.: Va et vient de l'iconicité en langue des signes française. In: Acquisition et interaction en langue étrangère, n°15 (2001) http://aile.revues.org/1405

Machines that listen: towards a machine listening model based on perceptual descriptors.

Marco Buongiorno Nardelli[1,2,3,4,5][0000−0003−0793−5055], Mitsuko Aramaki[5][0000−0001−6518−374X], Sølvi Ystad[5][0000−0001−9022−9690], and Richard Kronland-Martinet[5][0000−0002−7325−4920]

[1] CEMI, Center for Experimental Music and Intermedia, University of North Texas, Denton, TX 76203, USA
[2] iARTA, Initiative for Advanced Research in Technology and the Arts, University of North Texas, Denton, TX 76203, USA
[3] Department of Physics, University of North Texas, Denton, TX 76203, USA
[4] IMéRA - Institut d'études avancées d'Aix-Marseille Université, Marseille 13004, France
[5] CNRS, Aix Marseille University, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France
mbn@unt.edu
http://www.musicntwrk.com

Abstract. Understanding how humans use auditory cues to interpret their surroundings is a challenge in various fields, such as music information retrieval, computational musicology and sound modeling. The most common ways of exploring the links between signal properties and human perception are through different kinds of listening tests, such as categorization or dissimilarity evaluations. Although such tests have made it possible to point out perceptually relevant signal structures linked to specific sound categories, rather small sound corpora (100-200 sounds in a categorization protocol) can be tested this way. The number of subjects generally do not exceed 20-30, since it is also very time consuming for an experimenter to include too many subjects. In this study we wanted to test whether it is possible to evaluate larger sound corpora through machine learning models for automatic timbre characterization. A selection of 1800 sounds produced by either wooden or metallic objects were analyzed by a deep learning model that was either trained on a perceptually salient acoustic descriptor or on a signal descriptor based on the energy contents of the signal. A random selection of 180 sounds from the same corpus was tested perceptually and used to compare sound categories obtained from human evaluations with those obtained from the deep learning model. Results revealed that when the model was trained on the perceptually relevant acoustic descriptors it performed a classification that was very close to the results obtained in the listening test, which is a promising result suggesting that such models can be trained to perform perceptually coherent evaluations of sounds.

Keywords: Sound descriptors · Sound perception · Machine learning · Networks · Machine Listening

1 Introduction

Analyzing our surroundings through the many sounds that are continuously produced by our environment is a trivial task that humans do more or less automatically. Both natural sounds from the environment such as waves, rain, wind, or sounds from humans, machines or animals can be recognized and localized without any practicing. It is however much more complicated to tell a machine how to recognize such events through sounds. For that purpose we need to identify perceptually relevant sound structures for each source that somehow can be considered as the signature that characterizes an aspect of the sound, such as the action that caused the sound or the object, such as its shape, size or material of the sound source. Several previous studies have tempted to identify such characteristics and have identified sound structures linked to the perceived size [11] and the material of which it is composed [12,8,3]. In the case of more complex situations reflecting for instance interactions between sound sources, the listener perceives properties related to the event as a whole. Warren and Verbrugge [17] showed that objects that bounce and break can be distinguished by listeners with a high degree of accuracy, while Repp [14] revealed that subjects were able to recognize their own recorded clapping and the hand position from recordings when someone else is clapping. More recently, Thoret [16] showed that subjects were able to recognize biological motions and certain shapes from friction sounds produced when a person is drawing on a paper.

The present study focuses on how two different material categories, wood and metal, can be distinguished. Previous studies on the identification of material categories [4,3], have shown that both temporal aspects such as the damping and frequency related aspects such as the spectral bandwidth or the roughness are perceptually salient signal structures for such sounds. These approaches enabled us to design evocative sound synthesis models that enable to control sounds from verbal labels (material, size, shape etc). However, for more general uses, such as the identification of sound categories within large databases, the previous approaches are less adapted, since they rely on a combination of several acoustic descriptors obtained from a rather small set of sounds and therefore might not be adapted to general models that characterize environmental sounds. In the present study we therefore propose to focus on the log Mel energy of the sound by using a more global descriptor, namely the MFCC that has been commonly used in automatic classification tasks [9]. Although the MFCCs were initially designed for speech recognition based on source filter models [7], they integrate perceptual properties and mainly discard the source part making them rather pitch independent. It is therefore interesting to use such descriptors on large sets of sounds that cannot be easily pre-treated and equalized in pitch and intensity. Their ability to capture global spectral envelope properties is also an important advantage from a perceptual point of view [15].

The objective of this study is to take advantage of network-based modeling, analysis, and visualization techniques to perform automatic categorization tasks that mimic human perception. Similarly to social networks, gene interaction networks and other well-known real-world complex networks, the data-set of

sounds can be treated as a network structure, where each individual sound is represented by a node in a network, and a pair of nodes is connected by a link if the respective two sounds exhibit a certain level of similarity according to a specified quantitative measure. In this approach, one can see a lot of conceptual similarities between sound networks and social networks, where individual nodes are connected if they share a certain property or characteristic (i.e., sounds can be connected according to shared physical or perceptual properties, and people are connected according to their acquaintances, collaborations, common interests, etc.) Clearly, different properties of interest can determine whether a pair of nodes is connected; therefore, different networks connecting the same set of nodes can be generated.

In this paper we take a further step in exploring this promising direction of research. Specifically, many complex systems can be better analyzed via network representations as networks provide a nice mathematical tool to explore these systems. Uncovering the topological structures of networks may help to understand the organizing principles of underlying complex systems. Furthermore, the knowledge acquired via this approach has motivated us to explore machine learning models for automatic timbre characterization and classification and the effectiveness of such approaches compared to results from human perception tests. Automatic classification with deep learning approaches have previously been applied to large datasets of both speech and music [9,10], but fewer studies have investigated automatic timbre classification of environmental sounds.

The paper is organized as follows: in Sec. 2 we discuss the various methodologies employed in this research, from the definition of acoustical descriptors, to network metrics and machine learning models; in Sec. 3 we discuss the results of our analysis; we end with a few concluding remarks and a look towards future applications of this study.

2 Methodology

All the computational results in this paper have been obtained with the `MUSICNTWRK` package. `MUSICNTWRK` is a python library for pitch class set and rhythmic sequences classification and manipulation, the generation of networks in generalized music and sound spaces, deep learning algorithms for timbre recognition, and the sonification of arbitrary data. The software authored by one of the coauthors (MBN) and it is freely available under GPL 3.0 at www.musicntwrk.com [6][5].

2.1 Impact sound data.

We have compiled a database of ca. 1800 impact sounds produced by metal or wood objects from Splice Sounds collections. The length of each recording was equalized to 22050 sample points (5.0 sec. at a sample rate of 44100 Hz) by either zero-padding or truncation. Sounds span a broad palette of timbre and

provide a good data-set for statistical analysis. Of these 1800 sounds, we extract a random sub-set of 180 sounds that we used for the human perception tests and analyzed with network techniques. The remaining 1620 sounds have been used to train the machine learning models with a 80-20% split between training and validation sets.

2.2 Audio descriptors and metrics in the generalized timbre space.

Power Cepstrum and PSCC The Power Cepstrum of a signal gives the rate of change of the envelope of different spectrum bands and is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal:

$$\text{PSCC} = \left| FT^{-1} \left\{ \log(|FT\{f(t)\}|^2) \right| \right|^2 \tag{1}$$

In this work We always considered the first 13 cepstrum coefficients (PSCC), where the 0-th coefficient corresponds to the power distribution of the sound over time.

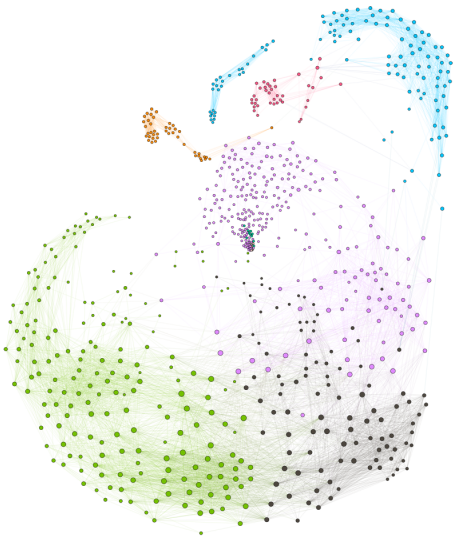


Fig. 1. Section of the network of the MFCCs built from the 1620 sounds that have been used to train the machine learning models. Colors indicate the classification based on their modularity class: Green, mostly high frequency tones from wood; Gray, mostly high to mid-frequency tones of wood; Purple, mostly deep frequency tones of wood; Cyan, mostly dry metal tones; Pink, mostly metal; and Orange, mostly "choked" metal sounds.

Mel Frequency Cepstrum and MFCC. The Mel Frequency Cepstrum of a signal is obtained as in Eq. 1 and differ from the power cepstrum by the choice of the spectrum bands that are mapped over the Mel scale using triangular overlapping windows. The mapping of the frequency bands on the Mel scale better approximates the human auditory system's response than the linearly-spaced frequency bands used in the normal cepstrum. We use a 16 bands Mel filter, and we considered the first 13 cepstrum coefficients (MFCC) as in the previous case. As for the PSCC, the 0-th coefficient corresponds to the power distribution of the sound over time.

Both PSCC and MFCC are obtained using 64 bins in the short time Fourier transform.

Networks and metric in timbre space. Network analysis methods exploit the use of graphs or networks as convenient tools for modeling relations in large data sets. If the elements of a data set are thought of as "nodes", then the emergence of pairwise relations between them, "edges", yields a network representation of the underlying set. Similarly to social networks, biological networks and other well-known real-world complex networks, entire data-set of sound structures can be treated as a network, where each individual descriptor (PSCC, MFCC) is represented by a node, and a pair of nodes is connected by a link if the respective two objects exhibit a certain level of similarity according to a specified quantitative metric. Pairwise similarity relations between nodes are thus defined through the introduction of a measure of "distance" in the network: a "metric". In this study we use the Euclidean norm (generalized Pythagoras theorem in N-dimensions) to quantify similarity between sound descriptors:

$$\text{distance}(I, J) = \sqrt{\sum_i \left(x_i^I - x_i^J\right)^2}, \tag{2}$$

where $\mathbf{x}$ is the chosen sound descriptor for sound I and J. In Figure 1 we show the network of the MFCC built from the 1620 sounds that have been used to train the machine learning models. In the figure we display the principal component for which less than 2% of all possible edges are built, that is, we allow an edge only if two MFCCs are at a distance that is less than 3% of the maximum diameter of the network. This representation reveals that the classification is coherent with material categories, as it can be observed from the emergence of clusters of sounds belonging to the same material (see for instance the green cluster of wood sounds on the lower left part of the network and the cyan, pink and orange clusters of metallic sounds in the upper part). This result validates the corpus with respect to the sound quality. For a general review on networks and graph theory the reader is referred to Albert and Barabási [2].

2.3 Machine learning model.

We implemented a deep learning model based on convolutional neural network (CNN) architecture inspired by similar approaches used in image and sound

recognition [13]. The CNN is built using the Keras kernel of Tensorflow [1] and it is trained on the full PSCC or MFCC data, after proper scaling and normalization. We retained only models with validation accuracy higher than 90%. After an appropriate model is chosen, it is tested on the set initially chosen for the human perception experiment. Each model chosen retains a similar accuracy on this set. A typical result of a training session on 30 epochs is shown in Fig. 2.

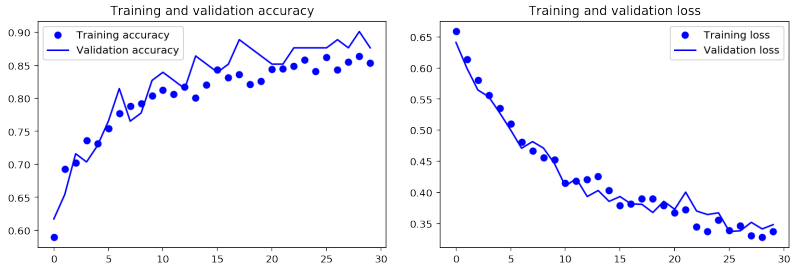


Fig. 2. Accuracy and loss in a typical model training run: (left) accuracy, (right) loss.

2.4 Experimental setup

The sounds were presented randomly to the participants thought headphones. The participants were asked to categorize each sound in either the Metal or Wood category by selecting the label shown on a graphical interface developed with Matlab. They could listen to each sound as often as desired.
Participants: Twenty-seven volunteers (21 males, mean age: 37 years-old) participated in the experiment. They declared no hearing nor cognitive problems.

3 Results and Discussion

A set of 180 impact sounds randomly extracted from the initial database (section 2.1) was used in a perceptual listening test. In Figure 3 and Table 1, we summarize the data of the perceptual test compared with four scenarios based on the machine learning models. In Figure 3 the scores represent the percentage of classification in the metal category. From this figure certain sounds (157) were clearly classified without ambiguity. 60 sounds were classified by 100% of the subjects in the labeled material category and 23 sounds that were less clearly classified were defined as "ambiguous", i.e. classified by less than 50% of the subjects in the labeled material category (for more details see Table 1). The above sounds were fed to our ML models in four different fashions: 1. with the model trained on the MFCCs to classify the sound with its MFCC (perceptual measure with perceptual model); 2. with the model trained on the MFCCs to classify the sound with its PSCC (physical measure with perceptual model); 3. with the

model trained on the PSCCs to classify the sound with its PSCC (physical measure on physical model); and 4. with the model trained on the PSCCs to classify the sound with its MFCC (perceptual measure on physical model). It should be noted that the ML models 1. and 3. have an accuracy that exceeds 90% in both cases. In the first column of Table 1 we report the percentage of classification in the metal category in the perceptual test. The successive columns list the sounds that are common with sounds characterized incorrectly by the four ML models (marked as X). Interestingly, we find a large degree of overlap between the perceptual scores and the ML scenario n. 4, physical model and perceptual measure. Moreover, a closer inspection to the data, shows that the majority of sounds, even if identified correctly by the model, retain a certain degree of uncertainty, as demonstrated by the probability of that sound to be characterized as metal listed in the last column. Indeed the majority of the scores are within $50\pm10\%$ and they could vary depending on the specific training of the machine learning model. To further support this observation, we have built the network of MFCCs for the full set of 180 sounds used in the perceptive test, shown in Fig. 4. The figure displays the first few giant components built with the shortest distances, less than 1% of all possible edges, that is, we allow an edge only if two MFCCs are at a distance that is less than 0.5% of the maximum diameter of the network. From the figure clusters of both unambiguous and ambiguous sounds can be observed confirming the observations made on the original network shown in Figure 1. It is evident that the "ambiguous" sounds are all clustered together and belong to the same modularity class demonstrating the robustness of the MFCCs as a relevant descriptor from a perceptual point of view.

4 Conclusion

By testing sounds using a perceptive measure (MFCC) on a machine learning model trained on physical parameters (PSCC), we reproduced a distribution of ambiguity in the classification of the origin of the sound that is coherent with the results of a human listening tests. In this way we can obtain scores that are coherent with perceptual tests. This is a first step towards a more general machine listening methodology that, if associated with perceptually salient acoustic descriptors that characterize the acoustic information used by the auditory system, might replace time-consuming listening tests and enable perceptual evaluation of huge databases of sounds. It could also be a valuable tool for cognitive studies to point out relevant sound structures (invariants) associated to perceptual categories based on very large data sets. Such sound structures open new possibilities to design evocative synthesis models that enable to control sounds in a perceptually consistent way.

Acknowledgments MBN wishes to acknowledge useful discussions with Alexander Veremyer and the financial support of IMéRA - Institut d'études avancées d'Aix-Marseille Université during his residency in Marseille in the Spring 2019.

Table 1. Table of sounds that are classified by less than 50% of subjects in the labeled material category in the perceptual test. We compare the perceptual scores with four different machine perception scenarios and list the sounds that overlap between the perceptual scores and the selected ML model. HM, percentage of classification in Metal category by human subjects (perceptual scores); MM, sounds in common with Model MFCC - MFCC; MP, sounds in common with model MFCC - PSCC; PP, sounds in common with model PSCC - PSCC; PM, sounds in common with model PSCC - MFCC; and ML, metal score (%) for model PSCC-MFCC. See test for a complete discussion

Ambiguous sounds	HM (%)	MM	MP	PP	PM	ML (%)
metal_092.wav	26				X	16
metal_106.wav	15				X	27
metal_119.wav	48	X				100
metal_1_04.wav	18					59
metal_1_12.wav	15					51
metal_2_06.wav	37				X	45
metal_2_10.wav	33					59
metal_3_10.wav	30					58
metal_4_03.wav	15					78
metal_4_10.wav	19					66
metal_1_06.wav	19					53
metal_3_03.wav	19				X	31
metal_4_05.wav	37				X	36
metal_4_11.wav	37				X	34
metal_1_02.wav	19				X	11
metal_1_03.wav	26				X	8
metal_2_04.wav	30				X	12
metal_3_07.wav	48				X	9
metal_3_12.wav	48				X	9
metal_1_11.wav	44				X	39
wood_019.wav	52		X		X	99
wood_005.wav	52					8
wood_16.wav	52		X		X	67

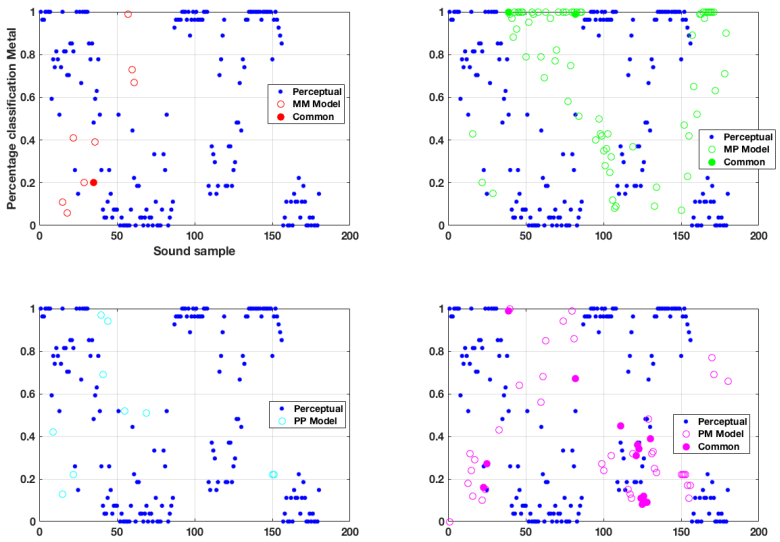


Fig. 3. In blue: Mean scores (corresponding to the percentage of classification in the Metal category) obtained from the perceptual test. These scores are compared with the scores of sounds characterized incorrectly by the four models (MM, MP, PM and MM models). The sounds that are in common are represented with filled markers.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (Jan 2002), https://link.aps.org/doi/10.1103/RevModPhys.74.47
3. Aramaki, M., Besson, M., Kronland-Martinet, R., Ystad, S.: Controlling the perceived material in an impact sound synthesizer. IEEE Transactions on Audio, Speech, and Language Processing 19(2), 301–314 (2011)
4. Aramaki, M., BRANCHERIAU, L., Kronland-Martinet, R., Ystad, S.: Perception of impacted materials: sound retrieval and synthesis control perspectives. In: Ystad, Kronland-Martinet, Jensen (eds.) Computer music modeling and retrieval: genesis of meaning in sound and music, pp. 134–146. Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2009), https://hal.archives-ouvertes.fr/hal-00462245
5. Buongiorno Nardelli, M.: `MUSICNTWRK`: data tools for music theory, analysis and composition. Proceedings of CMMR 2019 *in press* (2019), also at https://arxiv.org/abs/1906.01453

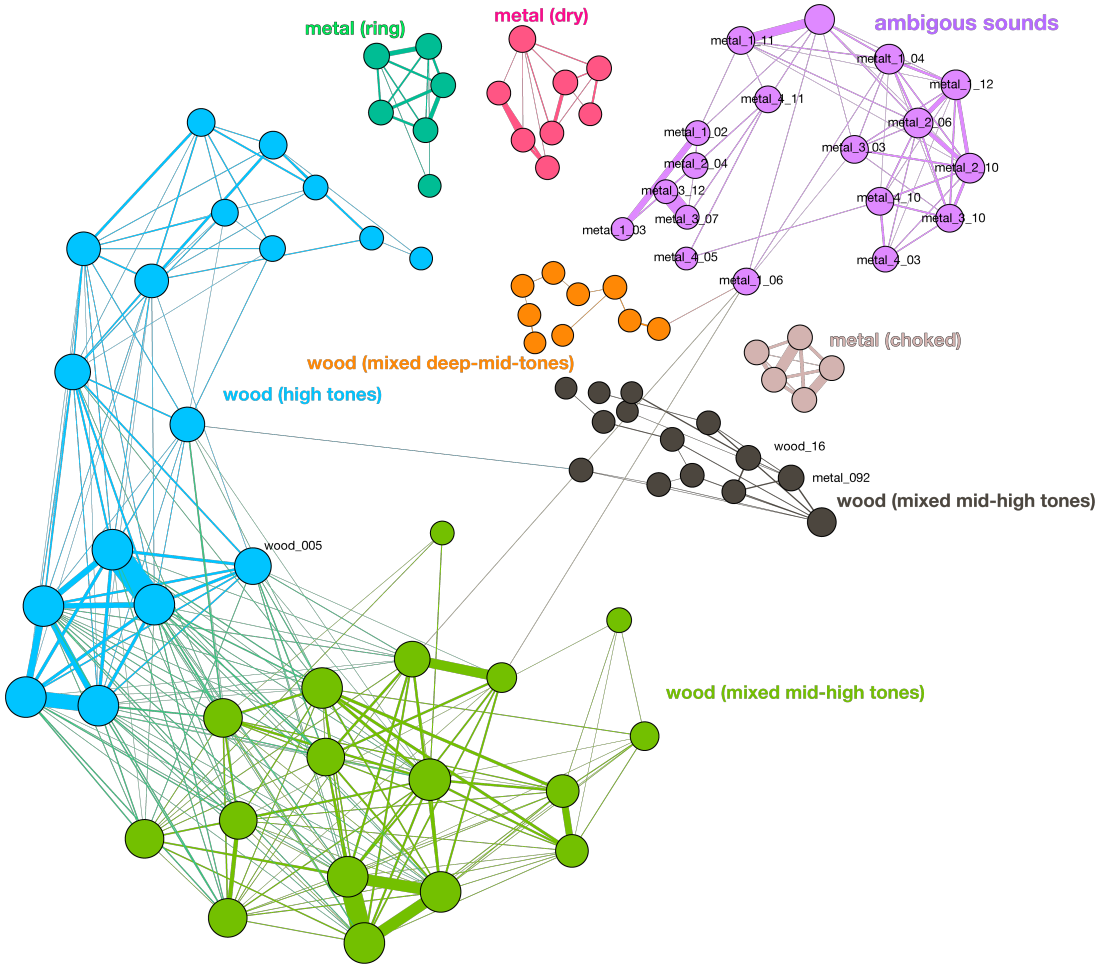


Fig. 4. Network of the MFCC of the sounds from the set used in the perceptual test.

6. Buongiorno Nardelli, M.: Topology of networks in generalized musical spaces. Leonardo Music Journal *in press* (2020), also at http://arxiv.org/abs/1905.01842
7. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing 28(4), 357–366 (August 1980)
8. Giordano, B.L., McAdams, S.: Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. The Journal of the Acoustical Society of America 119(2), 1171–1181 (2006)
9. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.W.: Cnn architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 131–135 (2017)
10. Kell, A.J., Yamins, D.L., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H.: A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 93(3), 630–644 (2018)
11. Lakatos, S., McAdams, S., Chaigne, A.: The representation of auditory source characteristics : simple geometric form. Perception and Psychophysics 59, 1180–1190 (1997)
12. McAdams, S., Chaigne, A., Roussarie, V.: Psychomechanics of simulated sound sources. Journal of Acoustical Society of America 115(3), 1306–1320 (2004)
13. Piczak, K.J.: Environmental sound classification with convolutional neural networks. 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP) pp. 1–6 (2015)
14. Repp, B.H.: The sound of two hands clapping: An exploratory study. The Journal of the Acoustical Society of America 81(4), 1100–1109 (1987)
15. Serizel, R., Bisot, V., Essid, S., Richard, G.: Acoustic Features for Environmental Sound Analysis. In: Virtanen, T., Plumbley, M.D., Ellis, D. (eds.) Computational Analysis of Sound Scenes and Events, pp. 71–101. Springer (2017), https://hal.archives-ouvertes.fr/hal-01575619
16. Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.L., Ystad, S.: From sound to shape: auditory perception of drawing movements. Journal of Experimental Psychology: Human Perception and Performance 40(3), 983–994 (Jan 2014), https://hal.archives-ouvertes.fr/hal-00939025
17. Warren, W.H., Verbrugge, R.R.: Auditory perception of breaking and bouncing events: a case study in ecological acoustics. Journal of Experimental Psychology : Human Perception and Performance 10(5), 704–712 (1984)

Pedaling technique enhancement: a comparison between auditive and visual feedbacks

Adrien Vidal,[1,2] Denis Bertin[1], Richard Kronland-Martinet[2], Christophe Bourdin[1]

[1] Aix Marseille Univ, CNRS, ISM, Marseille, France
[2] Aix Marseille Univ, CNRS, PRISM, Marseille, France
vidal@prism.cnrs.fr

Abstract. In cyclism, the pedaling technique is rarely optimal but could be improved using sensory feedbacks. The most common media used to display data of cycling power meter is a small screen placed on the handlebars. However, it could be dangerous by distracting the visual attention of the cyclist. That is why auditive feedback, called sonification, is investigated. In this paper, the effects of auditive or visual feedbacks on pedaling technique (evolution of the torque effectiveness) are compared using a lab experimental setup when subjects were engaged or not in a dual-task paradigm (cycling and detecting obstacles on the road). Improvement of pedaling technique is observed with both auditory and visual feedbacks, and reaction times to detect obstacles were not different between all conditions. However, sonification allows gaze behaviors more centered on the road, i.e. more secure. These results suggest that sonification could be a good solution to improve pedaling technique.

Keywords: Sonification, Gesture Efficiency, Cycling Ergometer, Torque Effectiveness, Cognitive Load

1 Introduction

Performance in cycling depends of a lot of parameters [1], [2], such as physiological factors, nutritional strategy, bike design, and also pedaling technique [3].

Technically, the pedal stroke simply consists in 4 phases: pushing and pulling phases and high and low transitions. Despite this apparent simplicity, the pedal stroke is rarely optimal even for expert cyclists, and difficulties are mainly observed during the pulling and transition phases, leading to loss of power then to less efficient performance. Based on this observation, the need to find efficient solutions to improve the performance has become a major issue in the domain of training but also of sport research.

The augmented reality approach, which consists in providing sensory information not naturally and directly available to the subjects remains one of the most promising technique. For instance, some studies have demonstrated that an augmented visual feedback may help to improve pedaling technique [4], [5]. However, for sports cognitively mastered with visually information like cycling (necessity to keep the eyes on the road), providing augmented visual feedback may overload cognitive processing [6] and distract vision from his major guiding role. For these reasons, augmented

auditory feedback (usually called "sonification" [7]), has been recently considered as a beneficial alternative for sport training [8], [9].

Our hypothesis is that auditive feedback could be a better solution compared to visual feedback to provide information to the cyclist by allowing him to keep the eyes on the road. However, these two modalities of sensory feedbacks in cycling have not been compared for now, and a comparison is conducted in this paper. Three characteristics are compared: the evolution of the pedaling technique, by means of torque effectiveness measurement, the gaze behavior, and the cognitive load induced by each sensory feedback. The paper is organized as follow: the second section of this paper details method of the experiment, the main results are presented in the third section, and results are then discussed in the fourth section. A conclusion ends the paper.

2 Method

24 participants took part to the experiment (mean age 26.2 +/- 9.4). They were not expert in cycling technique, but most of them practiced bike regularly. Three of the participants declared that the left foot was their preferred foot, the rest of participants declared to be right-footers. All participants signed an informed consent form in accordance with the Helsinki convention informing them about the conditions of the experiment and their right of withdrawal. The protocol was approved by the institutional review board of the Institute of Movement Sciences. The data were analyzed anonymously.

The experimental setup was made of a road bike Merida RaceLite, a HomeTrainer Tacx Flux and a screen (27 inches) placed in front of the cyclist, showing a virtual road moving in accordance to the cyclist cadence. The interface of this virtual environment was developed using the Unity platform and represented a straight road. The crank was a Rotorbike 2InPower, measuring the torque applied on both pedals. The Rotorbike crank was used with ANT+ transmission. The “fast-mode” of this crank was selected to transmit torque applied on each pedal at 50 Hz, whereas cadence and angular position were transmitted at 4 Hz. The resistance of the Home-Trainer was set to 130 W, and the cadence was not imposed: the cyclist had only to pedal at a regular and moderate cadence. Cyclist was also wearing Tobii Pro Glasses 2, allowing to measure the eyes behavior. A small screen (5 inches) was placed at the center of the handlebars, to display the visual feedback. Sounds were diffused through headphones (Sennheiser HD 201). Figure 1 represents the experimental setup.

Three conditions were assessed: one without any sensory feedback on performance (called “Control” condition hereafter), another using sonification (called “Auditive”) and a last one with a visual feedback (called “Visual”). For each condition, cyclist had to pedal during 3min. A rest (2min) was interleaved between each condition. The order of these conditions was permuted across all participants, in such a way all possible permutations were presented the same number of times.

These three conditions were presented twice: a first time without obstacles (reference task) on the road, and a second time with obstacles to be detected (dual-task). For a specific participant, order of conditions was the same with and without obstacles. For conditions with obstacles, the participants were required to detect them as quickly as possible by saying "Top". Oral answers were recorded, allowing to compute the Reaction Time (RT) for each obstacle detection. During each condition, 16 obstacles were presented, in four locations possible: Far away on the Left side (FL), Far away on the Right side (FR), Close on the Left side (CL), and Close on the Right side (CR). The instants of apparition of obstacles were randomly chosen but were the same for all participants. Two obstacles were separated by 3 s at least.

For the Auditive condition, a squeak was generated through headphones when the torque applied on the pedal was negative, lasting as long as the negative torque. The squeak was synthesized based on a Coulomb friction model, as presented by Thoret *et al* [10]. Both feet were sonified through a stereo reproduction: sound associated to the left (respectively right) pedal was diffused through the left (respectively right) earphone.

For the Visual condition, the same information was provided but visually on the small screen on the handlebars. A red light came on if the torque applied on the pedal was negative, lasting as long as the negative torque. As for the Auditive process, both feet were analyzed: a red light on the left side (respectively on the right side) of the screen was associate to the left pedal (respectively right).

Instructions were systematically read by the participants and then orally explained by the experimenter. These instructions presented the experiment and described the feedback process (the cyclist had to adapt his technique to avoid squeaking or turning on the red lights).

The data recorded with the crankset were sampled at regular time intervals, but this sampling did not enable each cycle or crank position to be analyzed independently of the speed rotation of the crank. Therefore, the torque was interpolated in such a way as to be sampled at regular angular intervals. The interpolation was performed in Matlab using the cubic interpolation of the function *interp1* with a 0.5° step. To assess performance on each stroke cycle, Torque Effectiveness TE was computed:

$$TE = 100 * \frac{T^{+} + T^{-}}{T^{+}}$$

with T^{+} the total positive torque over the cycle and T^{-} the total negative torque over the cycle (absolute value). TE was thus considered 100 % if there was no negative torque during the cycle. Mean Torque Effectiveness was computed for both feet (R-TE and L-TE).

Gaze behavior data recorded with the Tobii glasses were analyzed using the product software. 3 Areas of Interest were defined: the 27" screen displaying the road, the 5" screen displaying the visual feedback, and the rest of the visual field. Percentage of the time of visit duration was computed for each area.

Statistical analysis was conducted with Statistica. Four Repeated Measures Analysis of Variance (RM-ANOVA) were conducted:

- First, a two levels RM-ANOVA was conducted on the Cadence considering the factors Condition (three conditions) and Obstacle (with or without).
- A three levels RM-ANOVA was conducted on the Torque Effectiveness of both feet considering the factors Condition (three conditions), Obstacle (with or without) and Foot (left or right).
- A two levels RM-ANOVA was conducted on Reaction Time for obstacle detection considering the factors Condition (three conditions) and Position of the obstacle (four positions).
- A three levels RM-ANOVA was conducted on the percentage of the time of visit duration with the eyes considering the factors Condition (three conditions), Obstacle (with or without) and Area of Interest (three areas).

To go further these ANOVA, post-hoc tests applying the Bonferonni procedure were then conducted using a significance level of 0.05.

Figure 1: Experimental setup: 27" screen displaying a virtual road, 5" screen on the handlebars reporting visual feedback, bike Merida, Home-Trainer Tacx. The cyclist is wearing headphones during the Auditive condition, and is wearing Tobii glasses during the whole experiment.

3 Results

In this section, we analyze both the cadence, the torque effectiveness, the obstacle detection (RT and missed obstacles) and the gaze behavior.

3.1 Cadence

According to the ANOVA, the mean cadence did not significantly differ between Condition ($F(2, 46) = 0.206$, $p = 0.814$), and the presence of obstacles had no influence ($F(1, 23) = 3.747$, $p = 0.065$). The interaction of Obstacle and Condition also did not influence the cadence ($F(2, 46) = 2.009$, $p = 0.146$). The mean cadence for all conditions and all cyclists was 59.2 Rounds Per Minute.

3.2 Torque Effectiveness

The ANOVA yielded a significant effect of Foot ($F(1,23) = 4.382$, $p = 0.048$) and of Condition ($F(2,46) = 8.265$, $p = 0.001$), but no effect of Obstacle ($F(1,23) = 1.886$, $p = 0.183$). The Figure 2 reports the mean TE according to the Foot, and according to the Condition. Mean TE was significantly higher for the left foot (86.5 %) than for the right foot (85.2 %).

Mean TE during the Control condition (83.0 %) was significantly lower than the two other feedbacks (87.6 % during the Auditive condition, and 87.0 % during the Visual condition). However, there was no significant differences between the two conditions with sensory feedbacks.

The ANOVA did not yielded significant effect of interactions.

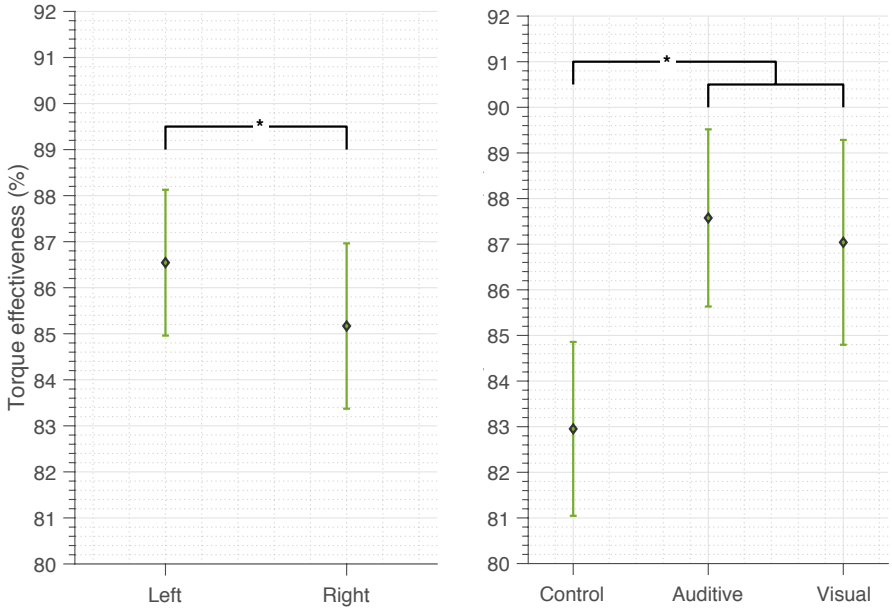


Figure 2: TE according to the foot (left) and according to the condition (right). Errorbars refer to a confidence interval of 95%. * means significant difference according to post-hoc tests with Bonferroni procedure, at significance level of 0.05.

3.3 Obstacle detection: Reaction Time and Missed obstacles

The ANOVA on RT yielded a significant effect of the Position of the obstacles ($F(3,69) = 6.916$, $p < 0.001$), no effect of the Condition ($F(2,46) = 1.214$, $p = 0.306$) and no effect of the interaction of the Position and the Condition ($F(6,138) = 1.622$, $p = 0.146$).

Figure 3 reports the mean RT according to the position of the Obstacle. Post-hoc tests revealed that obstacles at position CR (RT = 0.53 s) were detected quicker than obstacles at positions FL (RT = 0.61 s) and FR (RT = 0.59 s). RT of obstacles at position CL (RT = 0.56 s) was not significantly different from all others.

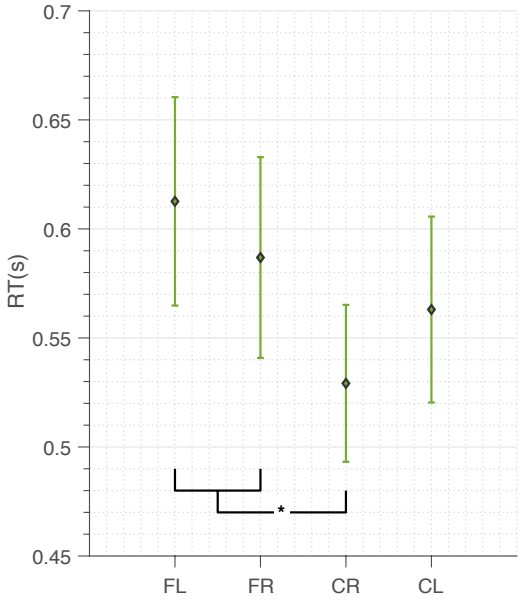


Figure 3: Reaction Time according to the position of obstacles (Front Left, Front Right, Close Right, Close Left). Errorbars refer to a confidence interval of 95%. * means significant difference according to post-hoc tests with Bonferroni procedure, at significance level of 0.05.

Figure 4 reports the missed obstacles according to the conditions. For each condition, there was a total of 384 obstacles to be detected (24 cyclists, 16 obstacles per condition). During the visual condition more obstacles were missed (6 obstacles) compared to other conditions. Only 3 obstacles were missed during the control condition and none during the Auditive condition.

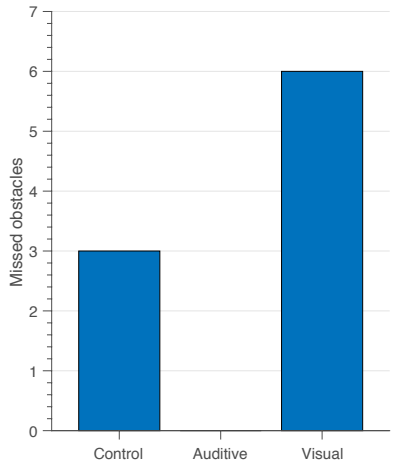


Figure 4: Total missed obstacles according to the condition

3.4 Gaze behavior

The ANOVA yielded a significant effect of the Area ($F(2, 30) = 106.61$, $p < 0.001$), a significant interaction of Obstacle and Area ($F(2,30) = 29.035$, $p < 0.001$), and a significant interaction of Condition and Area ($F(4,60) = 19.131$, $p = 0.001$).
Figure 5 represents the Total Visit Duration according to the Area of Interest. The 27" screen was significantly most viewed (73.5 %) than the other Areas (12.5 % for the 5" screen and 14.1 % for the rest).
The presence of obstacles on the road implied a modification of the visual behavior. Figure 6 represents the Total Visit Duration according to the presence of Obstacles and the Area of Interest. The 27" screen was significantly more looked when obstacles were present (97.3 %) than when obstacles were absent (50.0 %).
Figure 7 represents the Total Visit Duration according to the Condition and the Area of Interest. The Total Visit Duration of the 5" screen was significantly more important during the Visual condition (33.0 %) than during the two other conditions (1.1 % during the Auditive condition and 3.3 % during the Control condition). Moreover, the Total Visit Duration of the 5" screen was significantly less important during the Visual condition (56.2 %) than the two other conditions (79.8 % during the Control condition and 84.5 % during the Auditive condition).

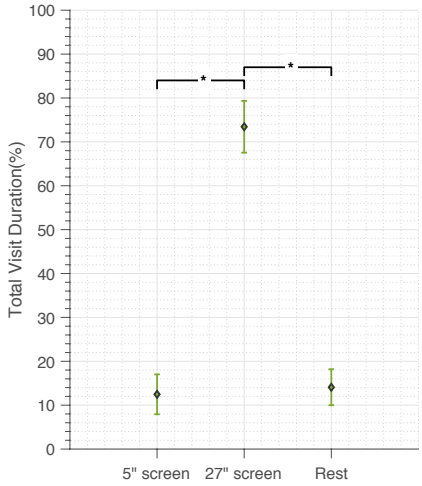


Figure 5: Total Visit Duration according to the Area of Interest. Errorbars refer to a confidence interval of 95%. * means significant difference according to post-hoc tests with Bonferroni procedure, at significance level of 0.05.

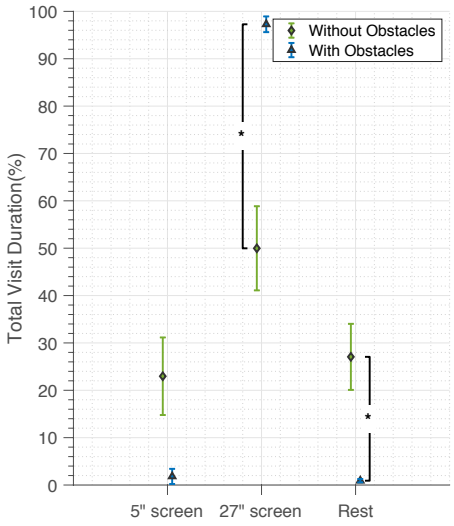


Figure 6: Total Visit Duration according to the presence of obstacles and the Area of Interest. Errorbars refer to a confidence interval of 95%. * means significant difference according to post-hoc tests with Bonferroni procedure, at significance level of 0.05.

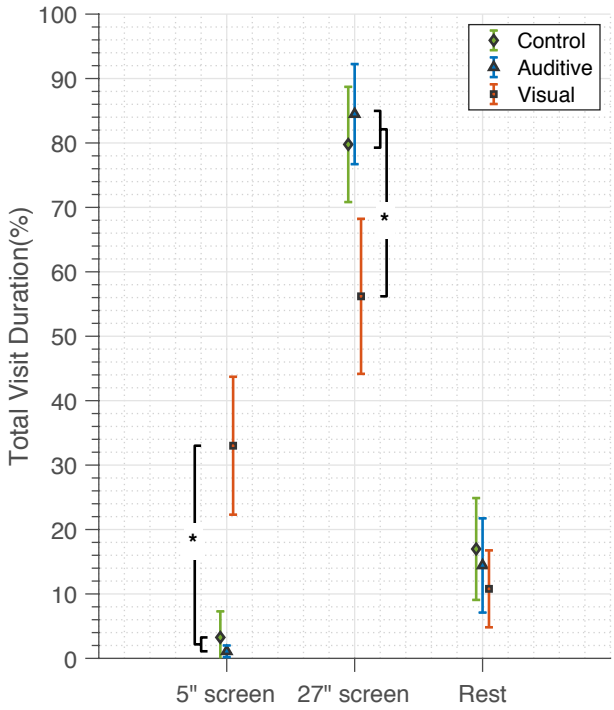


Figure 7: Total Visit Duration according to the Area of Interest and the Condition. Errorbars refer to a confidence interval of 95%. * means significant difference according to post-hoc tests with Bonferroni procedure, at significance level of 0.05.

4 Discussion

The goal of this experiment was to compare the effect of two modalities of sensory feedback used as augmented reality tools for the pedaling technique: auditive and visual feedbacks. These feedbacks were compared to a control condition (without any feedback). During the first part of the experiment, the cyclist had to focus only on his pedaling technique, whereas during the second part, some obstacles were placed on a virtual road and the cyclist had to detect them the most quickly possible.

Cyclists had to pedal at a regular cadence, but they had no specific information about their cadence (excepted the display of the virtual road moving according to the cadence). We first analyze the mean cadence of cyclists, and we showed that there was no significant effect of the feedback and the presence of obstacle on the cadence of

cyclists. The cadence could have an effect on the pedaling technique [3]. So, in this experiment, differences observed in pedaling technique were not linked to the cadence.

The TE is then analyzed, informing about the pedaling technique. TE was significantly higher for the left foot than for the right foot. An assumption explaining higher performances for the left foot is that cyclists could have a strongest leg [11], [12]. However, this assumption was not confirmed by the results, since there was no correlation between the observations and the dominant leg of cyclists, and few of the cyclists were left-footers. A further study should be conducted with half of left-footers participants.

TE was significantly lower during the control condition than during the two other conditions. It means that sensory feedbacks on instantaneous torque are effective to improve the pedaling technique, independently of the media.

During the second half of the experiment, cyclists had to detect obstacles on the road as quickly as possible, during the three conditions (Control, Auditive and Visual conditions). Reaction times were not different according to the conditions, suggesting that both feedbacks do not increase the cognitive load. However, some obstacles were missed during the visual condition, whereas none were missed during the Auditive condition.

Moreover, during the Visual condition, cyclists looked more often at the little screen on the handlebars, instead of looking at the road. This could in fact explain the targets missed during the visual condition. As attended, these results suggest that the gaze behavior could be dramatically impacted by the nature of the sensory feedback, visual cues leading the cyclists to take their eyes off the road to get information on their performance, whereas auditory cues allowing them to keep the eyes on the road. The Reaction Time measured in this experiment did not demonstrate this point, maybe because the task was too easy (obstacles were visible in peripheral vision).

Obstacles at position CR were the most quickly detected . It corresponded to the closest obstacles to the cyclists, that is the most dangerous for them. So this special area required the main attention of cyclist.

5 Conclusion

To enhance pedaling technique, visual and auditive feedbacks were compared in this paper during a two parts experiment: first, cyclists had only to focus on their pedaling technique, and in the second part they had to detect obstacles on a virtual road the most quickly possible while they concentrate on their pedaling technique.

Both feedbacks allowed significative enhancement of their pedaling technique in a similar way. However, during the visual condition, the gaze behavior was partially oriented towards the small screen on the handlebars presenting the augmented reality information, so that the cyclist was less attentive to its road (cyclists missed obstacles).

As a conclusion, our results confirm that, in an augmented reality approach, auditive feedback, said sonification, is a promising candidate to allowing for a significant improvement of the performance while preserving the security of the cyclists.

Then, to go further , a similar experiment have to be conducted in real conditions to definitively conclude on the promising effect of sonification on sport performance.

Acknowledgments. This research was funded by the French National Research Agency (ANR) under the SoniMove project (ANR-14-CE24-0018).

6 References

[1] G. Atkinson, R. Davison, A. Jeukendrup, et L. Passfield, « Science and cycling: current knowledge and future directions for research », *J. Sports Sci.*, vol. 21, nº 9, p. 767-787, sept. 2003.

[2] E. W. Faria, D. L. Parker, et I. E. Faria, « The science of cycling: factors affecting performance--Part 2 », *Sports Medicine*, 01-avr-2005. .

[3] R. P. Patterson et M. I. Moreno, « Bicycle pedalling forces as a function of pedalling rate and power output. », *Med. Sci. Sports Exerc.*, vol. 22, nº 4, p. 512-516, août 1990.

[4] D. Bibbo, S. Conforto, I. Bernabucci, M. Carli, M. Schmid, et T. D'Alessio, « Analysis of different image-based biofeedback models for improving cycling performances. », in *SPIE*, 2012.

[5] C. De Marchis, M. Schmid, D. Bibbo, A. M. Castronovo, T. D'Alessio, et S. Conforto, « Feedback of mechanical effectiveness induces adaptations in motor modules during cycling », *Front. Comput. Neurosci.*, vol. 7, 2013.

[6] R. Sigrist, G. Rauter, R. Riener, et P. Wolf, « Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review », *Psychon. Bull. Rev.*, vol. 20, nº 1, p. 21–53, 2013.

[7] S. Barras, « Auditory information design », PhD thesis, The Australian National University, 1997.

[8] L. Baudry, D. Leroy, R. Thouvarecq, et D. Chollet, « Auditory concurrent feedback benefits on the circle performed in gymnastics », *J. Sports Sci.*, vol. 24, nº 2, p. 149-156, févr. 2006.

[9] G. Dubus, « Evaluation of four models for the sonification of elite rowing », *J. Multimodal User Interfaces*, vol. 5, nº 3-4, p. 143-156, mai 2012.

[10] E. Thoret, M. Aramaki, C. Gondre, S. Ystad, et R. Kronland-Martinet, « Eluding the Physical Constraints in a Nonlinear Interaction Sound Synthesis Model for Gesture Guidance », *Appl. Sci.*, vol. 6, nº 7, p. 192, juill. 2016.

[11] F. P. Carpes, M. Rossato, C. B. Mota, et I. E. Faria, « Bilateral pedaling asymmetry during a simulated 40 km cycling time-trial », *Med. Sci. Sports Exerc.*, vol. 38, nº Supplement, p. S394, mai 2006.

[12] R. R. Bini et P. A. Hume, « Assessment of bilateral asymmetry in cycling using a commercial instrumented crank system and instrumented pedals », *Int. J. Sports Physiol. Perform.*, vol. 9, nº 5, p. 876-881, sept. 2014.

Musical Gestures: An Empirical Study Exploring Associations between Dynamically Changing Sound Parameters of Granular Synthesis with Hand Movements

Eirini- Chrysovalantou Meimaridou[1], George Athanasopoulos[2] and Emilios Cambouropoulos[3]

[1] Department of Music Studies, Aristotle University of Thessaloniki, Greece (eirini.meimaridou@gmail.com)
[2] Durham University, United Kingdom (georgios.athanasopoulos@durham.ac.uk)
[3] Department of Music Studies, Aristotle University of Thessaloniki, Greece (emilios@mus.auth.gr)

Abstract: This study explores the relationship between music and movement, focusing on hand movements in relation to electronically produced sound events (granular synthesis). This relation is studied empirically by presenting pairs of hand and sound gestures (in the form of videos) to participants, while trying to find cases where correlations exist between the two. More specifically, the focus is on properties of sound such as pitch, density and dispersion (in the context of granular synthesis), as well as on their association and description through hand gestures. A complementary goal is to examine whether any correlations exist between the properties of the hand movements (kinetic velocity, direction or surface) to the sound characteristics mentioned above. 48 participants (F: 29; R: 21-34) were asked to rate the goodness of fit between hand gestures and accompanying sound events. Participant responses confirm findings from previous studies, while new interesting observations, such as the connection between sound dispersion and kinetic energy of motion, are noted.

Keywords: Gestures, Musical Gestures, Embodied Music Cognition, Sound and Movement Correlations, Granular Synthesis.

1 Introduction

1.1 Musical Gestures - Music and Sound Associations

Recent years have seen an increased interest in investigating the association of music with non-musical concepts, especially with bodily motion. The human body "can be considered as the mediator between the person's environment and the person's subjective experience of that environment" [1, p. 5].

Undoubtedly, gestures often contribute to social interaction, facilitating communication and the attribution of meaning by means of hand movement [2]. During a musical experience, the human body interacts with music, and the human mind deals with the creation of interpretations related to this physical interaction [3]. Musical gestures involve the understanding of body, mind and environment, and their study is part of embodied music cognition [4]. Musical gestures may refer to a variety of pos-

sible actions with different functionalities: they may produce a sound, serve communication or, as it will be examined in this study, they may "describe", accompany or illustrate a sound event [2].

Music and movement are two concepts that interact and function supportively for each other; consequently, musical gestures are among the visible manifestations of this relationship. As a result, listening to a sound often leads people to creating correlations with concepts such as shape, material, size, direction, but also more abstract information such as colors or feelings [5].

People tend to associate various sound characteristics with physical space and movement, and consequently with musical gestures [6]. Sound features such as pitch and dynamics can be described as changes in the type, direction, or speed of a movement that accompanies them. Beyond that, gestures facilitate music understanding and music expression and are an important issue in the field of musical research [7].

On the one hand, the associations of music and movement rely on inherent tendencies and unconscious processes [8]. On the other hand, musical gestures may be influenced by numerous factors that emanate from a person's environment. "Although every music listener has a body, every culture constructs the human body differently" [8, p. 388]. Consequently, the performance of musical gestures is linked with social and cultural characteristics, such as language and music education [6], [8].

1.2 Related Work

As outlined in the previous section, people tend to "embody" sound and associate auditory stimuli with images, shapes or metaphors (kinetic or not). This section focuses on previous studies and findings that were the starting point for our research hypotheses.

Pitch is one of the most investigated music parameters in the context of sound and movement association. The most frequent and thoroughly examined correlation is the one concerning pitch and the vertical axis, and generally the notion of "height" [6], [9]. This is one relation between sound and movement examined in this study.

In addition, the present study explores the concept of constancy and its correlation with musical gestures. Here, the term 'constant' refers to an auditory stimulus that does not change during its evolution in time. According to the literature, the majority of people tend to associate time continuity with the horizontal x-axis [10] and with a left to right direction [11]. Another interesting finding is that the "description" of a sound with constant pitch, may guide people to the cessation of a producing action that accompanies the auditory stimulus [9]. The present study investigates how a gestural representation of a constant sound may be perceived when the x- axis is absent from the participants' choices.

As the auditory stimuli of this study concern sounds generated through granular synthesis, it was considered plausible to investigate some of the control parameters of this technique. Our interest turned to the concept of grain density and how it may be associated with (hand) gestures. Since the term tempo means the number of beats in a unit of time, and the term density means the number of grains in a unit of time, these concepts may show common trends, e.g. when associating sounds and motion. Tempo has been correlated with the concept of speed in previous research [6], so we hypothe-

size that density may also be described by this term. Additionally, we observed that two-dimensional graphic representations of grains (introduced by Ianis Xenakis), illustrate how different numbers of grains (and different density levels) can be represented in a delimited surface [12]. These findings prompted us to explore associations between density and speed movement or surface of movement. Since no previously published study has examined the possible relationship between sound dispersion (another control parameter of granular synthesis) and motion, we attempted to examine the correlation of dispersion with the term of speed and kinetic energy.

Finally, when people associate certain auditory stimuli with movement, the gestures they make evolve in time in a similar manner as the sounds do. This means that intensifying changes in sounds trigger corresponding kinetic intensifications, while musical abatements encourage motions with decreasing intensity [6].

1.3 Hypotheses

Based on the literature reviewed above, our hypotheses regarding the participants' responses are the following:

i. A constant sound (where density, dispersion and pitch of sound grains remain unaltered) will be associated with a motionless gesture.

ii. Modifications in grain density of the sonic stimuli will be linked to modifications in the surface of moving visual stimuli.

iii. Changes in grain dispersion of the sonic stimuli will be associated with changes of the hand gesture's kinetic energy and velocity of finger movement.

iv. Changes in pitch will be linked with changes of movement on the vertical axis

v. Opposite pairs of sound stimuli (increase and decrease in the density/ dispersion/ pitch) will be associated with opposite pairs of hand gestures (increase and decrease of movement's surface/ kinetic energy and velocity/ upward and downward gesture on y-axis).

2 Materials and Methods

2.1 Participants

The survey involved a random sample of people of diverse age, gender and musical knowledge background. In total, 48 people (F: 29; R: 21-34yrs, 27 self-identified musicians) participated in our study. Among the musicians, 23 participants were undergraduate or postgraduate university students majoring in music, while four were studying music at accredited conservatories for at least ten years. 14 musicians were trained as pianists, while the others were percussionists or string performers. The average duration of active music engagement was fourteen years.

2.2 Auditory Stimuli

Eight auditory stimuli (see Table 1) were synthesized using granular synthesis (Granulab VST 2 version)[1]. Among the granular synthesis parameters available, density, dispersion (grains within certain pitch range) and pitch were selected for further

[1] Granulab Inc. "Home". [Website] https://www.abc.se/~re/GranuLab/Granny.html

exploration. All stimuli were six seconds long and all alternations from the initial to the final state were linear and regular within this time span.

Table 1. Detailed Description of Sound Data

No	Sound Parameter	Description	Duration	Alterations		
				Density	**Dispersion**	**Pitch**
1	Constancy (zero dispersion)	Constant all parameters	6sec	10.5grains/sec	0 semitones	D6
2	Constancy (2-octave dispersion)	Constant all parameters	6sec	10.5grains/sec	2.4 octaves	random in range
3	Density	Increase	6sec	2.5→16 grains/sec	0 semitones	D6
4		Decrease	6sec	16→2.5 grains/sec	0 semitones	D6
5	Dispersion	Increase	6sec	10.5grains/sec	0 sem. → 2.4 oct.	random in range
6		Decrease	6sec	10.5grains/sec	2.4 oct. → 0 sem.	random in range
7	Pitch	Increase	6sec	10.5gr/sec	0 sem.	D6→E8
8		Decrease	6sec	10.5gr/sec	0 sem.	F#8→E6

2.3 Visual Stimuli

Eight videos were created as possible congruent visualizations of the sonic stimuli presented in the previous section, using recorded hand gestures. Each video lasted 6 seconds and was recorded using a Nikon D330 digital camera. The range of possible gestures was restricted to movements produced from the right hand's palm and fingers. In order to limit the endless variability of human movements, a forced- choice method of research was selected, narrowing participant choices to pre-recorded gesture representations.

The videos have been created so as to examine the main hypotheses explicated in the previous section. A secondary goal was to study gestures which can be recognized later by the hand gestural controller *Leap Motion*[2], so as to artistically exploit the results of the study at a later stage. Bearing in mind the limitations of the above technological device and, of course the type of movements we could associate with our auditory stimuli, we created gesture-videos characterized by four basic elements: i. lack of movement, ii. changes in kinetic energy and finger velocity (palm and finger movement), iii. changes in the surface of movement (palm opening/closing), iv. changes in the direction of movement along the y-axis. Each video is hypothesized to be congruent, with one of the sound conditions described in section 2.2, and incongruent with the other auditory stimuli (See Table 2).

[2] Leap Motion Inc. "Home". [Website] https://www.leapmotion.com/

Table 2. Visual Stimuli: Gesture's Description and Sound Mapping

Video	Gesture description	Congruent Sound
1	Palm and Fingers are open in the center of the screen - Motionless Gesture	Sound 1 – Constancy (zero dispersion)
2	Fingers moving continually with constant velocity - No changes in x or y axis	Sound 2 - Constancy (2-octave dispersion)
3 (See Fig. 1)	Palm and Fingers closed in the center of the screen - Gradual palm opening and increase of movement's surface - No changes in x or y axis	Sound 3 – Increasing Density
4	Palm and Fingers opened in the center of the screen - Gradual palm closing and decrease of movement's surface - No changes in x or y axis	Sound 4 – Decreasing Density
5	Palm and Fingers in the center of the screen - Initially still but gradually moving with increasing intensity - No changes in x or y axis	Sound 5 – Increasing Dispersion
6	Palm and Fingers in the center of the screen - Initially moving with decreasing velocity and finally reach immobility - No changes in x or y axis	Sound 6 – Decreasing Dispersion
7	Palm moving upwards on y axis	Sound 7 – Increasing Pitch
8	Palm moving downwards on y axis	Sound 8 – Decreasing Pitch

Fig. 1. Gesture no. 3 (Palm Opening) - frames at 0-3-6 seconds respectively

2.4 Procedure

The research was performed using a personal computer. The auditory stimuli were presented via headphones, in order to achieve optimal listening conditions and avoid distraction by external parameters. Initially, the participants had to listen to all eight sound tracks of the study. The aim was to get acquainted with "granular textures" and - perhaps - to observe the sounds and their different sonic characteristics per se (without involving the concept of movement). After this, the study was divided in three different parts.

The participants went through eight different pages/screens that contained the eight videos (visual stimuli) described in the previous paragraph. Each page, presented one of the eight different auditory stimuli (described earlier in paragraph 2.2) as a soundtrack to the videos; that is, all eight videos were presented in one page accompanied

by the same sound track. The participants were asked to see and listen to all the videos. In the first part of the study, listeners had to select up to three gestures that they considered to be the best representation of the sound. In the second part, the participants had to choose the most representative sound from the three they had selected. In the third part, participants encountered a screen/page that contained the all sounds and all videos (without sound) and they had to make a one-to-one mapping between them, i.e., they had to match every sound with a unique gesture representation (video). The participants had to fulfill these three tasks in fixed order, aiming to examine how they would respond to conditions that were gradually becoming more restrictive. Our aim was to investigate if different visualizations could potentially represent the same sound stimuli (as outlined in the 1st task), as well as to establish direct correlations between the gestures represented on the videos and the sonic stimuli when the possible choices were progressively limited by the participants themselves (through the 2nd task) and also on a one-to-one association (3rd task).

3 Results

The data analysis is divided into five different sections based on the hypotheses presented above (section 1.3). In total, 64 different combinations of sounds and gesture representations have been investigated (eight sounds * eight gesture videos). In this report we present results for all 48 participants as a single group of listeners. Initial examination of responses shows that both musicians and non-musicians selected similar gesture representations for the auditory stimuli – this analysis (including minor differences between the sub-groups) will be reported in a future publication.

In order to facilitate the presentation of data analysis, we use abbreviations for the hand gesture videos and for the different parts of the research procedure (1st part - Multiple Choices, 2nd part - Only One Choice, 3rd part- Matching). As the results of the 1st and 2nd part (and even the 3rd part) of the experiment are very similar, and due to restrictions of paper length, in the discussion below we will present the results for the first part of the study (Multiple Choices) and refer to the other parts only when necessary.

Table 3. Code names for hand gestures

Code Name	Description	Code Name	Description
[P=]	Palm still	[F<]	Still to moving fingers
[F=]	Moving fingers	[F>]	Moving to still fingers
[P<]	Palm opening	[U]	Moving up
[P>]	Palm closing	[D]	Moving down

3.1 Constant sound

Choosing the most representative gesture for the sound which is a constantly repeating pitch at a constant grain density, seems to have prompted participants to a variety of different responses as seen in Figure 2a.

Specifically, at the first part where the participants could choose up to three possible answers, most of them associated this sound with the gesture [U] that shows the hand moving upwards on the vertical axis (25.97%); they associated also this sound

with the gesture [P=] that shows a still gesture (23.37%) but also with the gesture [P<] where the hand palm is gradually opening (22.07%). Similar associations are given at the second part of the experiment. At the final part of the experiment, however, 87.5% of the participants associated the constant sound with the still gesture [P=].

The choice of the still gesture indicates that they tend to associate a sound that does not change during its duration with a hand movement that also does not change. On the contrary, the association with changing gestures [P<] and [U], maybe shows the tendency of listeners to represent the evolution of time that is the only changing parameter here. It is also noteworthy that gestures [F =], [F <], [F>] were selected in a small percentage (or not at all).

3.2 Changing Density

In Figure 2c, we observe a clear preference of the participants to correlate the increasing density sound with the gesture indicating palm opening [P <]. The next most prevalent answer is the [F <] video (increasing kinetic energy and speed of fingers). The response rates are as follows: at the first stage, the gesture [P<] involves 40.5% of the participants' answers and gesture [F<] involves 22.78%.

The participants associated the sound that is characterized by decreasing density mostly with the gesture closing palm gesture [P>]. More specifically, at the first stage, the gesture [P>] involves 36.66% of the participants' answers and gesture [F>] involves 22.22%.

3.3 Constant/ Changing Dispersion

When associating a sound with constant dispersion with a gesture representation, the majority of the participants chose to correlate this auditory stimuli with the video [F =] with response rates 58.82% at the first part of the experiment – see Figure 2b. The random scattering of grains within a certain range of frequencies was associated with the gesture that depicts the fingers of the hand moving constantly.

Participants' answers gave a fairly clear image concerning the association of a sound characterized by increasing dispersion with a gesture. Most of them chose gesture [F<] that shows the hand's kinetic energy and the fingers' speed increasing as the most representative (51.42%). But, also, 22.85% of the answers concerned the [F=] gesture that involves moving fingers at a constant rate. The association of sound characterized by decreasing dispersion, led the participants to select the video [F>] (40.27%) and we also observed a preference for gesture [F=] (26.83%).

3.4 Changing Pitch

The participants' answers when associating an increasing pitch sound with a gesture showed clear preference for the gestures [P<] and [U]. Specifically, at the first stage 43.15% chose the [P<] video and 34.73% the [U] video. At the final stage (one-to-one matching), 66.66% showed a clear preference for the gesture [U].

Finally, when associating a sound with decreasing pitch, the participants chose the opposite gesture representations. 26% chose the gesture [P>] and 31.95% the [D]. At the final part, most of the participants chose the [D] representation (60.41%).

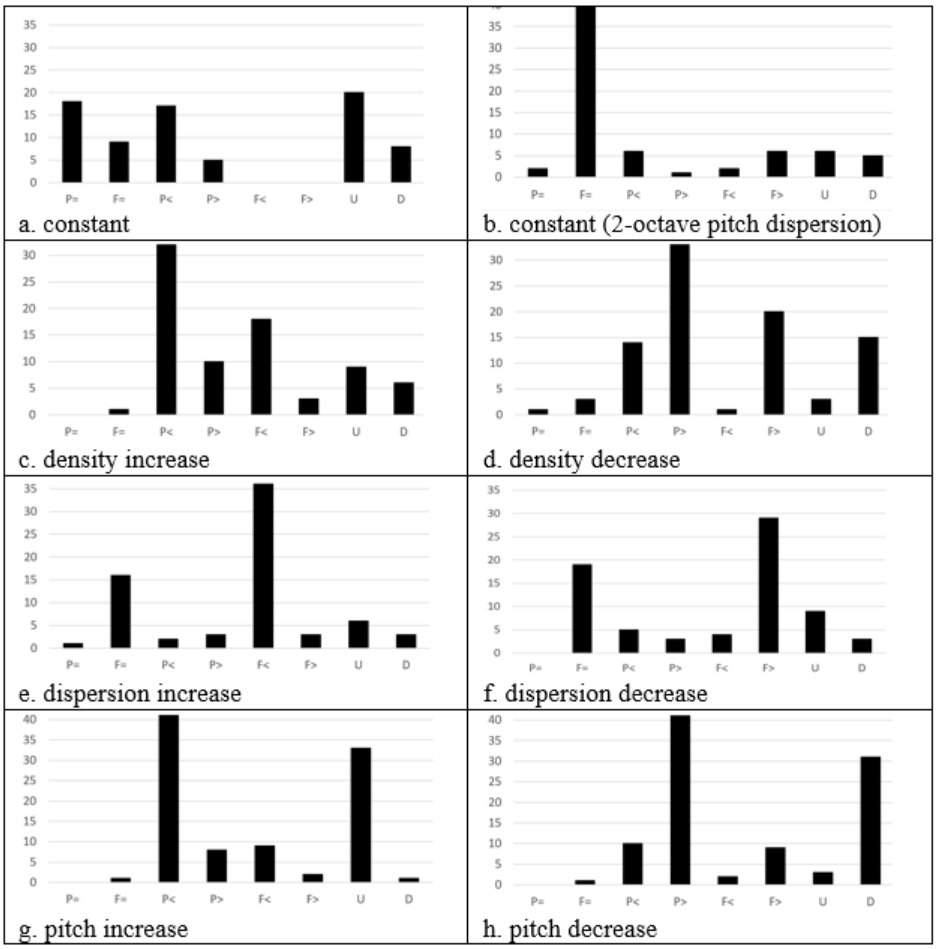


Fig. 2. - Bar diagrams showing number of times each of the eight gestures (P=, F=, P<, P>, F<, F>, U, D) have been associated with each of the eight sound stimuli during the 1st stage of the experiment.

3.5 Opposite Sound Pairs

The participants associated opposite pairs of auditory stimuli (characterized by increasing/ decreasing density, dispersion and pitch), with pairs of gestures exhibiting a reverse movement (increase and decrease of movement's surface/ kinetic energy and fingers' velocity/ upward and downward gesture on y-axis). In the majority of cases the number of responses was similar between the opposite correlations. In some cases, small asymmetries and pronounced tendencies for one of the two opposite states (ascending or descending) occurred. The following table illustrates these relations.

Table 4. Total amount of participant' responses concerning opposite sound pairs

	Associated Gesture	**1st stage - Multiple Answers**	
		Total	**Percentage %**
1st sound pair (density)	[P<]	32	40.5
	[P>]	33	36.66
	[F<]	18	22.78
	[F>]	20	22.22
2nd sound pair (dispersion)	[F<]	36	51.42
	[F>]	29	40.27
3rd sound pair (pitch)	[P<]	41	43.15
	[P>]	41	42.26
	[U]	33	34.73
	[D]	31	31.95

4 Discussion

The present study followed a forced- choice methodology. Participants were asked to associate musical stimuli with hand gestures and were not allowed to make their own gestures to visualize the sound stimuli. Should participants had produced gestures that embody sounds in a free-representational manner (see, for example [11], this would have resulted in obtaining a too large variety of gestures on pre-existing and well-established norms (e.g., pitch in relation to vertical height time in relation to horizontal axis). As such, this approach would not enable us to delve into more subtle, yet discernible approaches in gesture representation (e.g., movement of fingers). Participants had to associate auditory stimuli that consisted of diverse sound characteristics (1. constant/ changing density, 2. constant/ changing dispersion, 3. changing pitch) with gesture representations that also manifested manifold characteristics (1. lack of movement, 2. alternations in the surface of movement, 3. kinetic energy and fingers' velocity, 4. direction of movement). When we asked 48 participants to combine the concepts of sound and movement, and to discover how different changes that concern sounds made in the context of granular synthesis could be represented via hand gestures, many of our initial hypotheses have been confirmed. At the same time, new and interesting findings became apparent.

4.1 Constancy

Concerning the constant sound (track no. 1), the participants did not agree on the type of gesture that would represent this auditory stimulus. There was no strong correlation between this sound and what was hypothesized to be the congruent visual stimuli (i.e. static image), but, instead, various different gestures were associated with this sound. The association of the constant sound with the motionless gesture only partly confirms our initial hypothesis.

The association of constant sound with the two other options (the gradual opening of the hand palm [P <] and the upward movement on the y-axis [U]) underlines the participants' inclination to represent the evolution of time. Although people tend to

represent time on the x- axis [10], the lack of this choice in the visual stimuli has prompted participants to other matches (opening palm and y-axis).

4.2 Density

The sound and gesture associations that concerned the auditory stimuli with changes in the grain density (tracks no. 3 and 4) showed common trends for listeners in all the parts of the study: participants preferred the opening palm gesture [P <] as more suitable for a sound that is characterized by increasing density and closing palm gesture [P>] for decreasing density. There is, therefore, a tendency to correlate the increase of the sound grain with the increase in the surface area of the movement and vice versa.

4.3 Dispersion

Regarding the sound characterized by constant dispersion (track no.2), the results show, as hypothesized, that it was mostly associated by participants in all three parts of the study, with the gesture [F =]. That is, listeners correlated this sound with a gesture that is characterized by a continuous motion of the fingers at constant velocity.

The participants' choices concerning the sounds characterized by changing dispersion (tracks no. 5 and 6), showed a clear preference for associating these with the gestures characterized by a change in kinetic energy. The sound with a gradual increase in dispersion was associated with gesture [F <] (gradually increasing kinetic energy of fingers), and, vice versa. The next option was gesture [F =] that involves constant moving of fingers. In any case, it seems that listeners tend to represent changes in sound dispersion with intense kinetic activity of the fingers (static/increasing/decreasing).

4.4 Pitch

Changes in pitch were associated with changes of movement on the vertical axis, as hypothesized, but also with modifications to the surface of movement. Participants' answers showed that they tend to associate changes in pitch with changes in the direction of movement on the vertical axis. Thus, a sound characterized by increasing pitch was represented by a gesture showing the hand ascending on the y-axis, while a sound of decreasing pitch was represented by a gesture where the hand descends along the same (vertical) axis. This finding (correlation of pitch to the vertical axis) has been explored and confirmed by the majority of researchers in the last fifty years [6], [9].

Additionally, a high number of the participants associated the increase in pitch with a gesture of an opening palm (increase of the surface of movement) and vice versa (for a sound characterized by pitch decrease). This conclusion matches Antovic's findings concerning the correlation between pitch and size concept [13]. The latter observed that when people describe sounds with metaphors, there is a tendency to translate high pitch sounds into large-scale shapes, while sounds of low pitch were associated with small-scale shapes.

4.5 Opposite Sound Pairs

Finally, the results showed that the predicted association of opposite sound couples with opposite gestures has been confirmed. Listeners showed a tendency to match opposite auditory stimuli pairs with opposite gesture couples.

Pairs of sounds characterized by increase and decrease in density, dispersion, and pitch were identified by the participants and have been associated with gestures exhibiting contrasting intensity and direction (increase and decrease of movement's surface/ kinetic energy and fingers' velocity/ upward and downward gesture on y-axis). This finding is in line with previous studies exploring the existence of proportions when associating opposite auditory pairs and gestures, such as the research carried out in 2006 by Eitan and Granot [6].

5 Conclusion and Future Directions

In the present study, we investigated gestural representations of different sound parameters of granular synthesis, including density, dispersion and pitch. Constant and changing conditions of the parameters mentioned above were selected, and we examined if they were associated with specific hand movements. The principal scope of this study has been to assess how familiar sensorimotor experiences correspond to sonic stimuli in a forced-choice design, using Granulab as a tool for developing the sonic stimuli. Though studies have been conducted on artificial/tangible musical instruments as compositional tools using granular synthesis mappings [14], in this case our aim was to focus primarily on putting pre-existing associations between gesture and sonic events to the test, and not the compositional approach in itself. By offering a three task approach to the participants (primary selection of up to three gestural representations to each stimulus; narrowing this selection to one gesture per stimulus; free association of gestures to stimuli) our aim to assess the association between a specific set of gestures to sounding responses has been achieved.

The results confirmed most of our hypotheses and correspond to the findings of previous studies. At the same time, new interesting outcomes emerged and enriched our knowledge about sound and motion correlations. Absence of sonic manipulations (constant sound) has been primarily associated, as hypothesized, with lack of movement; we observed, however, an increase in overall hand surface and an upward movement on the y-axis which we conjecture corresponds to the 'temporal movement' of a sound event. Changes in sonic density have been associated mostly with changes in the movement's surface, while changes in dispersion have been associated with changes in the kinetic energy and finger velocity. The concept of pitch was closely correlated with the direction along the vertical axis (as expected) but also with the surface of movement (palm opening and closing). Finally, the participants tended to correlate opposite pairs of auditory stimuli with opposite gesture couples.

Future research may examine different sound synthesis parameters and well as movement gestures in different combinations, and may move beyond the forced-choice methodology used in this study to use gestural controller and/or motion capture techniques to track freely shaped gestures by participants exposed to different sound stimuli.

In conclusion, an artistic and research blend as a future scope of research would be a great challenge and a logical continuation of the current study. Using a gestural controller (e.g. Leap Motion Controller) and appropriate software (e.g. Pure Data or Max/MSP) could easily transform the study's results and the sound-gesture mappings into an interesting artistic interactive music project.

6 References

[1] M. Leman, "Musical gestures and embodied cognition," *Actes des Journées d'Informatique Musicale (JIM)*, Mons, Belgium, pp. 5–7, 2012.

[2] A. R. Jensenius, M. M. Wanderley, R. I. Godøy, and M. Leman, "Musical Gestures: Concepts and Methods in Research," in *Musical Gestures: Sound, Movement, and Meaning*, New York: Routledge, 2010, pp. 12–35.

[3] M. Leman, *Embodied music cognition and mediation technology*. Cambridge: MIT Press, 2008.

[4] R. I. Godøy and M. Leman, "Why Study Musical Gestures?," in *Musical Gestures: Sound, Movement, and Meaning*, New York: Routledge, 2010, pp. 3–11.

[5] A. R. Jensenius and H. T. Zeiner- Henriksen, *Music Moves: Why does music make you move?* Massive Open Online Course [MOOC]. University of Oslo, 2017.

[6] Z. Eitan and R. Y. Granot, "How Music Moves: Musical Parameters and Listeners' Images of Motion," *Music Percept.*, vol. 23, no. 3, pp. 221–247, 2006.

[7] A. Gritten and E. King, Eds., *New perspectives on music and gesture*. Aldershot, United Kingdom: Ashgate, 2011.

[8] V. Iyer, "Embodied Mind, Situated Cognition, and Expressive Microtiming in African-American Music," *Music Percept.*, vol. 19, no. 3, pp. 387–414, 2002.

[9] M. B. Küssner and D. Leech-Wilkinson, "Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm," *Psychol. Music*, vol. 42, no. 3, pp. 448–469, 2014.

[10] D. Casasanto and K. Jasmin, "The hands of time: Temporal gesture in English speakers," *Cogn. Linguist.*, vol. 23, no. 4, pp. 643–674, 2012.

[11] M. B. Küssner, "Shape, Drawing and Gesture: Cross- Modal Mappings of Sound and Music," King's College London, 2014.

[12] T. Lotis and T. Diamantopoulos, *Music Informatics and Computer Music.* Athens: Hellenic Academic Libraries, 2015 (In Greek)

[13] M. Antovic, "Musical Metaphors in Serbian and Romani Children – an Empirical Study," *Metaphor Symb.*, vol. 24, no. 3, pp. 184–202, 2009.

[14] G. Essl and S. O'Modhrain, "An enactive approach to the design of new tangible musical instruments," *Organised Sound*, vol. 11, no. 3, pp. 285–296, 2006.

Concatenative Synthesis Applied to Rhythm

Francisco Monteiro[1], F. Amílcar Cardoso[2], Pedro Martins[2], and Fernando Perdigão[1]

[1] DEEC, University of Coimbra
fmonteiro@alunos.deec.uc.pt; fp@deec.uc.pt
[2] CISUC, Department of Informatics Engineering, University of Coimbra
{amilcar,pjmm}@dei.uc.pt

Abstract. Music and technology have a long symbiotic history together and became even more related in recent times, when technology is so present in our everyday lives. There is an ongoing need to develop new music creation tools that explore the possibilities presented by technological breakthroughs. Different forms of sound synthesis have surfaced over the years, including concatenative synthesis, which offers the ability to create music using large databases that modern computers can handle. We propose a concatenative synthesizer oriented towards rhythmic composition, which has the ability to create its own drum kit by interpreting an input drum audio signal. The system then recreates the input through different ways, including a variation algorithm based on Euclidean rhythms. It was implemented in the programming language Max/MSP and the extension Max For Live, in order to make it usable in the DAW environment. We have also created a basic interface to interact with the user.

Keywords: Concatenative Synthesis · Rhythm Composition · Audio Effect · Audio Plug-In · Max For Live.

1 Introduction

Algorithmic composition is the set of techniques, languages or tools aiming to compose music in an entirely automated way. Its application has been in constant parallel development with technological breakthroughs and different models have risen ever since the creation of the field [6]. Throughout the years, multiple models [10,4] and musical pieces were created and along with the computational capacity increase, more complex and attractive solutions have risen too.

In the past decade we have seen an increased interest in machine learning models and that also meant an increased interest in algorithmic composition. Developments in the music information retrieval field have allowed new possibilities [12], including developments in concatenative synthesis as well as drum transcription [17].

Concatenative synthesis can be defined as a form of sound synthesis that uses a database of sound segments called *units*, and a unit selection algorithm to find sequences of units that match an objective, called the *target* [14]. Originally

used in speech synthesis [7], it then saw its first musical applications in the early 2000s. The artistic context presents different objectives, meaning a different set of criteria was used to find new solutions.

This paper presents an audio effect that seamlessly recreates an input drum phrase automatically and intelligently, including coherent rhythmic variations according to the controls of the user. This involves a system that is capable of translating the sound input into a sequence of meaningful drum units and reuse those units as the sound source for a rhythmic recreation.

2 Related Work

Inspired by the speech synthesis approach to concatenative synthesis, Schwarz [13] applied the same principle to music creation in his Caterpillar system. The use of the Viterbi algorithm by his speech synthesis peers was not beneficial to music creation because the artistic domain imposed a new objective: finding coherent units and transitions but also variating solutions, which resulted, ultimately, in a different approach.

A constraint-satisfaction mechanism or *music mosaicing* is proposed in [18] as a solution to the limitation imposed by the Viterbi determinism. Instead, it uses a heuristic algorithm to find an "acceptable" solution, as opposed to the "ideal" solution of the Viterbi algorithm.

Both these past implementations lacked real-time user interaction. In order to counter that, in 2006 Schwarz [15] introduced *CataRT*, which included an *interactive timbre space*. The corpus is mapped in a 2-dimensional space according to two user selected parameters.

Bernardes, Guedes, and Pennycook [1] combined the user interface of Schwarz with Tristan Jehan's *Skeleton* [8] — a creative system that accounted for the perceptual listening models of the human hearing system — to create *EarGram*, a software designed for PureData which introduced new visualization features along with clustering possibilities.

Guedes and Sioros [5] proposed a real-time rhythm generation algorithm based on a stochastic model that automatically creates "generic" rhythms (they do not belong to a specific musical style) from a certain meter and metrical subdivision level provided by the user.

Also in the rhythm composition domain, Cárthach Ó Nuanáin [11] introduced *RhythmCAT*, a VST Plug-In that uses a Concatenative Synthesis approach for rhythm generation. Given a *seed* rhythmic loop, the plug-in segments up the different individual instruments in the sequence through host tempo information, extracts features (loudness, spectral centroid, spectral flatness and MFCCs) from each of the *seed* units and then matches the individual units to a user selected and previously analyzed corpus of units. The seed sequence is then recreated with the corpus units with respect to a controllable concatenation cost, calculated by the weighted Euclidean Distance of the extracted features. The interface also includes the 2-dimensional map of previous examples.

3 Description

We propose a plug-in that receives a drum loop and returns a reorganization of the individual instruments into a new sequence and introduces algorithmic variations, dependant on the user's decisions. We implemented it using the Max/MSP language and its extension for the Ableton Live environment: Max For Live. During the description of the system, we will use the following nomenclature:

Step Smallest time unit defined and corresponds to the time length of a 16^{th} note.
Segment Smallest audio unit defined and corresponds to the audio content within the length of a 16^{th} note.
Phrase Audio content within the length of a bar.
Event Segment that contains an audio onset — Figure 1.
Sample Largest group of Segments in between two consecutive events — Figure 2.
List or Message The equivalent to both a string and a vector in Max/MSP language, although in this work is best to think about it always as a numeric vector. We will use the terms interchangeably.
Collection A group of Messages/Lists in Max/MSP.

Its operation cycle has got the length of one bar in a fixed 4/4 signature. We only work with relative time lengths, the absolute time lengths are defined by the BPM chosen in the host application — Ableton Live in this case. Each unit is a drum sample that was segmented from the input audio source.

For the purpose of sampling and segmentation, we use *chucker~*[3], a MAX/MSP object that given an audio input, a division measurement — 16 in our case — and a synchronization signal — provided by our Master and Host Ableton Live —- will divide the phrase into an equal number of isochronous segments. This fixed length segmentation is a limitation, as it prevents us to work with groove and general temporal fluctuations present in drum recordings that are important to their rhythmic quality. We will further discuss this limitation in the conclusions. Thankfully, audio stretching algorithms are capable of artificially synchronizing onsets to fixed time markers — see Fig. 1.

chucker~ operates with a bar of delay because it needs to fill its data buffer with audio content. After the first bar, chucker~ is fully operational. To inform *chucker~* about the desired output sequence, we send it a list —- the MAX/MSP version of a vector: a message/string composed of spaced numbers. We will refer to this list as *Y*. This list has 16 elements, in a range 1-16. The index of the list states the step's temporal positioning, while the actual value of each element corresponds to the output segment itself S(x) — see Fig. 1. The following two examples illustrate this:

$$\begin{aligned} \text{Y(original)} &= [1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15\ 16] \\ \text{Y(reverse)} &= [16\ 15\ 14\ 13\ 12\ 11\ 10\ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1] \end{aligned}$$

[3] https://docs.cycling74.com/max7/refpages/chucker~

Fig. 1. Time domain visual representation of the first bar of a drum break in *Melvin Bliss – Synthetic Substitution* in Ableton Live environment. Quantized through audio stretching to every 16^{th} note. The markers are called Warp Markers and denote transients that were time shifted — in our nomenclature these are the *events* of the drum break.

chucker~ also requires a Boolean list with 16 elements, stating witch steps are active and which ones are silent — list D. Active steps are true and silent steps are false.
To have a complete example, these would be our *chucker~* arguments if we wanted to loop the first 4 segments of the original phrase only during the first half of the phrase — Y values are irrelevant when their step is inactive in D:

$$\begin{aligned} \mathrm{Y} &= [1\ 2\ 3\ 4\ 1\ 2\ 3\ 4\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0] \\ \mathrm{D} &= [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0] \end{aligned}$$

There are four distinct *chucker~* instances running at the same time in order to make it possible to have polyphony: one *chucker~* associated with each drum sound and an extra one to run a variation algorithm.

3.1 Analysis

The analysis of the segments is a crucial part of the plug-in, since it will create the criteria that makes the Concatenative Synthesis coherent — see Figure 3. For onset detection, we used *aubioOnset~* in high frequency content mode, which is an accurate onset detector for percussion [3]. For the transcription, we used a device created by Marcus Miron et al. called Live Drum Transcription. It implements an instance-filtering method using sub-band onset detection [9] to create a drum classifier with live applications. The device receives an input audio signal and outputs bangs out of three outlets which state the occurrence of an event in each of the three classes available: *Kick*, *Snare* and *Hi-Hat*. We run Ableton's clock inside Max/MSP to translate these pulses into quantized step values.

The lists with the segments that are associated with each class — events — are stored in a collection called *DrumEvents*. Whenever Ableton enters Play

Fig. 2. First bar of a drum break from *Aerosmith - Walk This Way* to illustrate what can go wrong with fixed length segmentation. Between the first 4 segments, only S1 contains an onset — is an event —, although the drum lasts at least 3 more steps. We associate the event with its length. In this case the length is 4 segments because the next event is located in S5. So, the sample 1 includes the segments S1, S2, S3, S4.

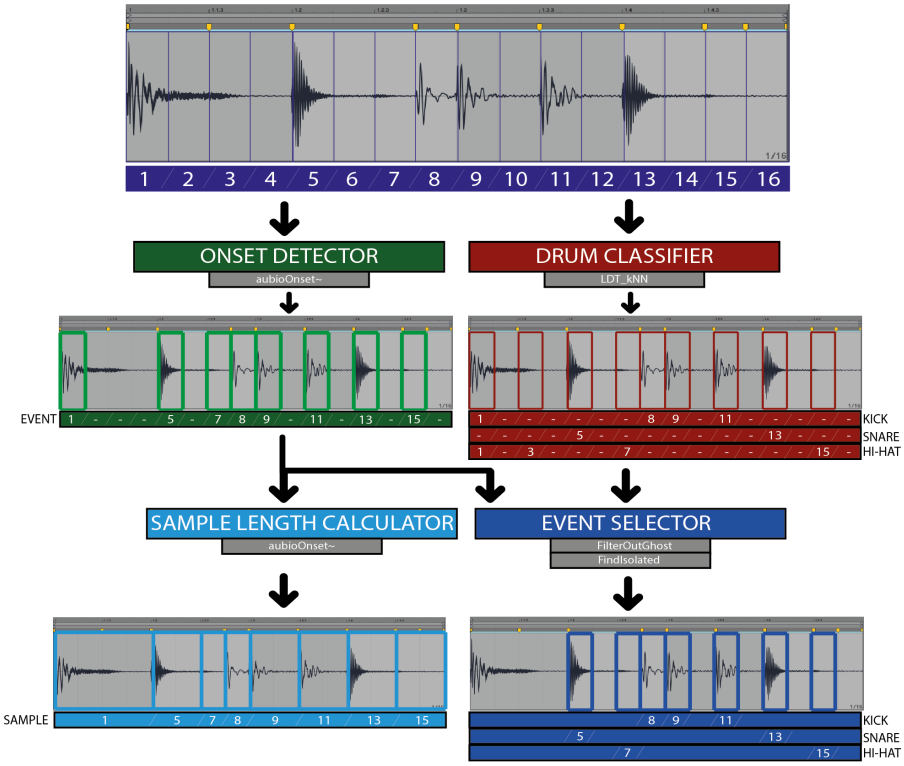


Fig. 3. Diagram of the whole Analysis Process using the same file as Fig. 2. Through this figure, we can see how the program creates its database/drum kit from the input drum phrase.

Mode, the analysis starts too. LDT is inconsistent and sometimes the output varies, and for that reason, analysis will run until there are 4 consecutive bars that output the same event list for each class. When this happens, the program enters *Stability* mode: analysis stops and we process *Analysis*.

LDT can sometimes output ghost events, particularly in the Hi-hat class, so we also store a list of segments that include an onset — *Events*. We then match *Events* lists with *DrumEvents* list and eliminate the ghost events in *DrumEvents* — segments that were transcribed as a drum sound but were not detected by *audioOnset* — and store the remainder in another collection called *EventFiltered*.

In drum phrases, it is common to have multiple instruments being played at the same time, and LDT can classify a segment as multiple instruments. Obviously, it is beneficial to work with isolated drums if possible, so the plug-in tests which events are only associated with a single class. In case our drum phrase does not include isolated drums for a particular class, we maintain the previous *EventFiltered* results.´

One of the problems with using *chucker~* to segment the phrase is that it segments in fixed intervals, independently of the sound characteristics. As a consequence it may happen to cut a *Sample* that lasts longer than the period of a segment — see Figure 2 — and because LDT is mainly based on onset detection, the segments that trigger a detection are only the ones that contain an onset. The plug-in corrects this by measuring a distance between consecutive entries of the *Events* list. This way, when we intend to play a sample, we can play all the segments that are involved in it in sequence.

3.2 Synthesis

Like stated before, we use four different *chucker~* instances to manage each of our concatenative voices. The first three *chucker~* instances will be associated with the *Main Sequence*, a single bar structure that will serve as the base loop over which we will layer our variation algorithm.

We offer the possibility to reproduce MIDI files. We created a MATLAB script to convert MIDI files into *.txt* files stored in the plug-in folder, that can be interpreted by Max/MSP as collections.

The other *chucker~* is meant to be layered with the *Main Sequence*. In order to generate new notes, we will have a random process running with the periodicity of one bar. This process generates notes according to the Euclidean Rhythms.

Euclidean Rhythms are a family of rhythms computed through the Euclidean algorithm that has been used in mathematics to compute the greatest common divisor given two integers. Its ability to generate evenly spaced patterns has seen applications in string theory, computer science and nuclear physics. This same algorithm can be used to elaborate a series of rhythmical ostinatos that describe rhythmical patterns found in various traditional music cultures all over the world. Their relationship to music was first discovered by Godfried Toussaint and is described in [16].

The formulation of the rhythmical sequences is based on the computer algorithm invented by Eric Bjorklund [2]. He used a binary representation of the problem, which can also be used in rhythmical description. Each bit describes an equal time interval, 1s represent the occurrence of an event — the onset of a note - and 0s represent the absence of events — a rest. A common and short way of representing these binary sequences is *E(m, k,o)*, with m being the length of that sequence, k the number of events in it and o the offset variable which will circularly shift the whole sequence by an integer value:

$$E(10,4,0) = [1\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 0]$$
$$E(10,4,1) = [0\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 1]$$

Since we are now working with samples that may contain more than one segment, we need a new form of rhythmic representation. We will call it *O(x)*, and its a binary sequence that states the occurrence of events — not to be confused with *D(x)* that states where there is reproduction of audio content and silence. All our values in this subsection will be based on the analysis example in Figure 3.

Besides rhythmic variation, we included another type of variation in our concatenative synthesizer, which deals with unit selection. Given that we already know which steps will include events, both in the *Main Sequences* and in the *Euclidean Rhythm*, and represent that in the form of the binary sequence *O(x)*, we can now handle the sample assignment. Our selection is simply random. If we have multiple samples classified as the same instrument, when we need to reproduce that particular instrument we will randomly assign one of the classified samples. Knowing that out output sequence O_{SNARE} contains snares in steps 3; 6; 10; 13 and the input phrase contains snare events in segments 5, 13, if we use the unit decision variation, one alternative could be:

$$\mathrm{O}_{SNARE} = [\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0]$$
$$\mathrm{Y}_{SNARE} = [\ 0\ 0\ 5\ 0\ 0\ 13\ 0\ 0\ 0\ 13\ 0\ 0\ 5\ 0\ 0\ 0]$$

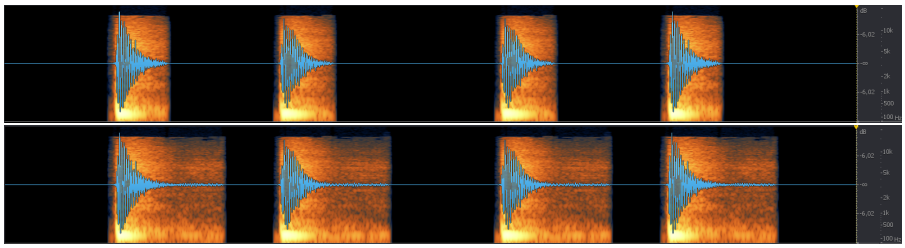


Fig. 4. Spectral representation of the mentioned Snare sequence. The upper spectrum is the sequence containing only the events associated with the snare sounds. The lower spectrum concatenates the event with the remaining segment that is part of the sample.

But since we now know that the samples associated with events 5 and 13 contain more than one segment — two segments each — we have to correct the argument

messages in order to make *chucker~* reproduce the whole sample lengths. See Figure 4 for an illustration of this.

$$\mathrm{Y}_{SNARE} = [\ 0\ 0\ 5\ 6\ 0\ 13\ 14\ 0\ 0\ 13\ 14\ 0\ 5\ 6\ 0\ 0]$$
$$\mathrm{D}_{SNARE} = [\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0]$$

Even though we use a single *chucker~* for each instrument to achieve polyphony, a single *chucker~* is only capable of reproducing one audio stream. It could happen that a particular sample would overlap with another, but our system does not allow it. We do so by prioritizing drum events over full sample playback, because with percussion the most important part of the sample is always the onset.
Going back to our previous sequence as an example. If we have an extra snare event on step 2 and it gets assigned the sample 13 from unit selection which has a sample length of 2 steps and overlaps with a posterior event — on step 3 — we cut its length to the maximum possible length that prevents overlapping:

$$\mathrm{O}_{SNARE} = [\ 0\ \mathbf{1}\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0]$$
$$\mathrm{Y}_{SNARE} = [\ 0\ \mathbf{5}\ 5\ 6\ 0\ 13\ 14\ 0\ 0\ 13\ 14\ 0\ 5\ 6\ 0\ 0]$$
$$\mathrm{D}_{SNARE} = [\ 0\ \mathbf{1}\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0]$$

3.3 Interface

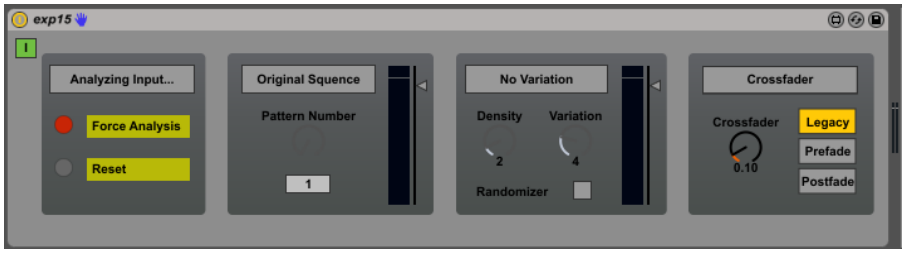


Fig. 5. Max For Live integrated interface in Ableton environment.

In Figure 5 we present our front-end of our software. We will explain the function of each control:

Force Analysis Forces the analysis stoppage and begin the concatenation process with the current analysis parameters.
Reset Resets the analysis process from the beginning. Useful in case the plug-in made a wring analysis or if the user wants to change the input drum phrase.
Original Sequence / MIDI files Button that toggles the Main Sequence option.

Pattern Number Selects which MIDI file should be reproduced in the output. The knob automatically sets its range according to the number of converted MIDI files in the saved text files.

No Variation / Euclidean Algorithm Button that toggles the Variation option.

Density Minimum value of events in the Euclidean algorithm.

Variation Range of possible events in the Euclidean algorithm.

Randomizer Allow the assignment of unfiltered events in the Variation.

Crossfader Control the crossfader amount in between segments.

Legacy / Prefade / Postfade Assign the crossfade mode of all *chucker~* objects.

4 Tests and Evaluation

One of the main problems imposed by this application was to find a testing method that would retrieve meaningful results. We decided to take the tool to the testers in a computer already set up properly with the plug-in, with drum phrases that we knew would work and with MIDI converted files. We tested it with 10 different individuals with different relations with music. All of them had experience with interacting with music, not just listening, but either playing instruments, broadcasting radio shows, playing records as DJs or were actually involved in music production/composition. They were showcased a little demonstration, then proceeded to experiment the plug-in themselves and were finally asked to reply to a series of questions through the *Google Forms* platform:

1. On a scale 1-5, classify the utility of this tool in a music production context:
2. On a scale 1-5, classify the utility of this tool as a percussive accompaniment for to practice an instrument:
3. On a scale 1-5, classify the usability of the tool:
4. On a scale 1-5, classify the quality of the variation generator:
5. What other functionality would you like to see in the plug-in?
6. What is your relationship with music:

The results were fairly positive in most questions. The variation generator received a lot of good comments as sounding "genuine" or "rhythmically coherent" — see Figure 6.

There was a couple of negative comments regarding unit segmentation, which we had previously diagnosed. That is mostly a problem of *chucker~* and its fixed segmentation grid, even with the time-stretching Quantization. The crossfader options are too limited and even with some tweaks on the crossfader section, it never really sounds as good as a regular sampler with an ADSR (Attack-Decay-Sustain-Release) type of envelope.

Another comment we received regarded the length of the loops. Since we only had a 1 bar loop running, it could be a good addition to be able to use longer drum loops so it would not be so repetitive. That is actually just a problem of

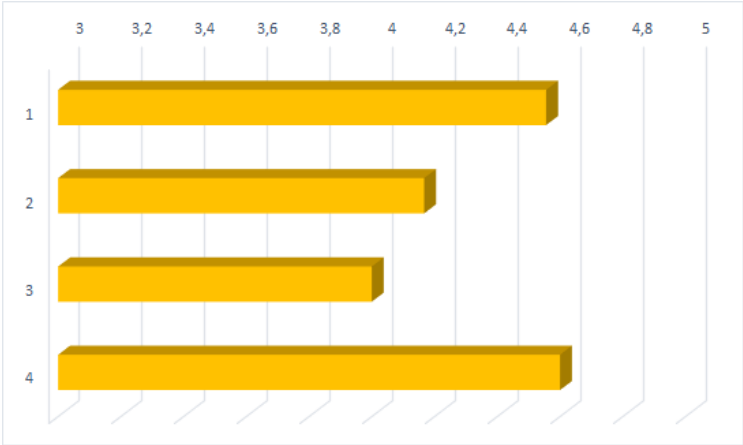


Fig. 6. Average Score for each of the first 4 questions.

our implementation, and again using *chucker~* as our main engine complicates many of these ideas, even though we also considered them.

Finally, another criticism we received regarded the inability to pause the program and maintain the analysis. That was also a problem of our implementation, but possibly this one could be avoided. It was something we implemented from the very beginning, as we invested quite some time in the analysis section, and it became convenient to restart the analysis by pausing and playing in the debugging stage. Eventually, it became a central part of the program and would be time-consuming to fix it, so we decided to keep it that way and focus on other more urgent and fracturing problems.

5 Conclusions and Future Work

This paper presents a functional plug-In that would be useful in the electronic music creation context. We created a concatenative synthesizer focused on rhythmic-centric loop recreation. The environment chosen was the Max/MSP programming language for the back-end development, and the extension Max For Live for the DAW Ableton Live as, respectively, user-interface and host program. Our system is able to receive a drum phrase — using traditional ocidental drum instruments: Kick, Snare, Hi-Hat —, segment that drum phrase into individual samples and resequence those drum samples into new drum phrases. These new drum phrases can be selected from the conversion of MIDI files, for which we provide a MATLAB script. The Plug-In includes a functional variation algorithm that applies Euclidean geometry concepts to, additively, modify the drum phrases.

Various users were asked to experiment the plug-In, evaluate it and comment on it, which resulted in positive feedback. Besides having a few fixed questions

for every user, we also had some productive dialogs regarding the utility and the future of the plug-in.

We made this sort of implementation to work in live applications, but this kind of analysis would make sense in many systems. Ableton Live, for example, when it automatically segments a drum phrase it does not label each segment like our program does. This can be time-saving for people who rely on this technology on an everyday basis.

Ultimately, it sounds musical. And the individual samples sound good together. The database is created by the input drum phrase and, most the time, these drum phrases were recorded in a properly tuned drum kit and played by a professional drummer. Even if we change the sample sequence, they come from the same source. That is probably also a reason why this kind of implementation works.

We recognize that a lot of what was done in Max/MSP is not the best way to implement that particular solution. We will address some possible improvements to the plug-in. Some of them were not added because they would not fit the plug-in in its current state, but we are open to reformulate some of its features:

chucker~ Our main sampler started being used as a prototyping tool, but soon it became central to the system and it presented various limitations, such as the fixed segmentation grid, the inability to express groove and velocity. So we would definitely rethink the segmentation process. This also opens up the possibility for expressive simulation, velocity variation, unlimited bar length, multiple envelope modes, different time signatures and many other possibilities that most sequencers already contain.

MATLAB script It is very simple and can be definitely done in JavaScript, which MAX/MSP supports, meaning everything could be done in the Max For Live interface.

Classification We have used the LDT classifier, which works good enough for what we wanted to achieve here, but if it was possible to have even more classes and more accuracy, we could widen the palette of drum phrases the system could work with.

Offline MIDI device Taking the ideas implemented here, we could implement a similar MIDI device that segments audio and automatically labels each segment.

Rhythm Generation Although we have created a plug-in that is able to create rhythmic variation, we do not consider it to be generative. There are many database-oriented Machine Learning approaches to rhythmic generation that would make sense to include in a system of this kind, like analyzing a database of MIDI files from a genre and create an generate new drum phrases of that genre, or analyzing different drum solos from a particular drummer and create an automatic improvisation tool that mimics his/her style.

Melodic Content As melodic detectors become more accurate, we can envision a plug-in that recreates previously composed music available in MIDI with samples translated from other audio sources.

References

1. Bernardes, G., Guedes, C., Pennycook, B.: Eargram: an application for interactive exploration of concatenative sound synthesis in pure data. In: Lecture Notes in Computer Science, From Sounds to Music and Emotions, Revised Selected Papers of the 9th International Symposium on Computer Music Modelling and Retrieval. vol. 7900, pp. 110–129 (2013)
2. Bjorklund, E.: The theory of rep-rate pattern generation in the sns timing system. SNS ASD Tech Note, SNS-NOTE-CNTRL-99 (2003)
3. Brossier, P.M.: Automatic annotation of musical audio for interactive applications. Ph.D. thesis (2006)
4. Fernández, J.D., Vico, F.: Ai methods in algorithmic composition: A comprehensive survey. Journal of Artificial Intelligence Research **48**, 513–582 (2013)
5. Guedes, C., Sioros, G.: A formal approach for high-level automatic rhythm generation. In: Proceedings of the BRIDGES 2011 (2011)
6. Hiller, L.: Composing with computers: A progress report. Computer Music Journal **5**(4), 7–21 (1981)
7. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. vol. 1, pp. 373–376 vol. 1 (May 1996). https://doi.org/10.1109/ICASSP.1996.541110
8. Jehan, T.: Creating Music by Listening. Ph.D. thesis, MIT (2005)
9. Miron, M., Davies, M.E., Gouyon, F.: An open-source drum transcription system for pure data and max msp. In: Proc. of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 221–225 (2013)
10. Nierhaus, G.: Algorithmic composition: paradigms of automated music generation. Springer Science & Business Media (2009)
11. Ó Nuanáin, C.: Connecting Time and Timbre : Computational Methods for Generative Rhythmic Loops in Symbolic and Signal Domains. Ph.D. thesis, UPF (2017)
12. Schedl, M., Gómez, E., Urbano, J., et al.: Music information retrieval: Recent developments and applications. Foundations and Trends® in Information Retrieval **8**(2-3), 127–261 (2014)
13. Schwarz, D.: A System for Data-Driven Concatenative Sound Synthesis. In: Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00). pp. 97–102 (2000)
14. Schwarz, D.: Current Research in Concatenative Sound Synthesis. In: International Computer Music Conference (ICMC) (2005)
15. Schwarz, D., Beller, G., Verbrugghe, B., Britton, S.: Real-Time Corpus-Based Concatenative Synthesis with CataRT. In: Proc. of the 9th International Conference on Digital Audio Effects (DAFx). pp. 279–282. Montreal, Canada (Sep 2006)
16. Toussaint, G.T., et al.: The euclidean algorithm generates traditional musical rhythms. In: Proceedings of BRIDGES: Mathematical Connections in Art, Music and Science. pp. 47–56 (2005)
17. Wu, C.W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Muller, M., Lerch, A.: A Review of Automatic Drum Transcription. IEEE/ACM Transactions on Audio Speech and Language Processing **26**(9), 1457–1483 (2018). https://doi.org/10.1109/TASLP.2018.2830113
18. Zils, A., Pachet, F.: Musical mosaicing. Proceedings of the COST G-6 Conference on Digital Audio Effects (DaFx-01) pp. 39–44 (2001)

Enhancing Vocal Melody Transcription with Auxiliary Accompaniment Information

Junyan Jiang[1,2,3], Wei Li[2], and Gus G. Xia[3]

[1] School of Computer Science, Carnegie Mellon University
[2] Computer Science Department, Fudan University
[3] Computer Science Department, New York University Shanghai
junyanj@cs.cmu.edu, weili-fudan@fudan.edu.cn, gxia@nyu.edu

Abstract. Monophonic vocal melody transcription is a classic task of content-based music information retrieval. However, for note-level transcription, the performance of current systems still cannot meet practical requirements. In this paper, we propose a novel algorithm to improve the performance of note-level monophonic vocal melody transcription by combining a Convolutional Recurrent Neural Network (CRNN) a the Conditional Random Field (CRF). Moreover, we propose a context-assisted method with two kinds of widely available auxiliary information, i.e., accompaniment audio and approximate word-level lyric timestamps. Experimental results show that our system significantly outperforms the baseline, and auxiliary accompaniment information can further improve the performance even when it is partially provided.

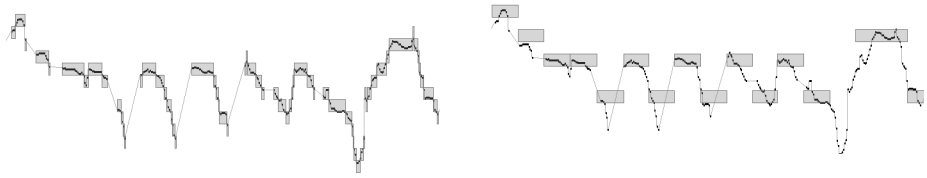
Keywords: Singing melody transcription, deep learning, auxiliary information, context-assisted method

1 Introduction

1.1 Monophonic Vocal Melody Transcription

Note-level vocal melody transcription is useful for many Music Information Retrieval (MIR) applications, such as query by humming [6] and singing evaluation [18]. To acquire the note-level representation of a singing melody, traditional vocal melody extraction systems mostly adopted a two-stage method [5, 12, 19]. First, the continuous fundamental frequency (f_0) contour is extracted from the audio. Then, this continuous contour is used to decode the discrete note sequence. While the first stage is generally considered a solved problem for monophonic cases with robust fundamental frequency trackers such as PYin [14], the performance of the second stage is still far from satisfactory.

The f_0 contour of a vocal track contains stylistic features such as vibratos and sliding notes [11]. These features hinder the decoding of note sequence as they introduce noisy frequency deviations that often lead to onset detection errors and semitone errors when quantizing f_0 curves to discrete note sequences. An example of such errors is shown in Fig. 1.



(a) (b)

Fig. 1: An illustration of pitch quantization error from f_0 curves by comparing (a) nearest pitch quantization and (b) ground-truth labeling. Excessive notes (*gray blocks*) are decoded due to large fluctuation of f_0 curves (*black curves*) caused by singers' stylistic features such as sliding notes.

Some systems tried to model the stylistic features of f_0 contour in order to increase the transcription accuracy. For example, The Tony system [12] allows a larger frequency deviation around note onsets and much smaller deviation for the sustain (stable) part of a note. Yang et al. [19] considered the character of vibratos and created a pitch dynamic model. However, the stylistic features of f_0 vary from singer to singer, and they are generally hard to be modeled by rules and traditional machine learning methods.

Deep learning models have recently been widely adopted for MIR tasks [4]. To better model the stylistic features, we introduce the Convolutional Recurrent Neural Network (CRNN), a powerful architecture that has been widely used in computer audition to model complex acoustic features from speech [1], music [15] and general sound [2], and apply it to our task of note-level monophonic vocal melody transcription.

1.2 Auxiliary Accompaniment Information

In practice, we can often acquire useful auxiliary information of the vocal melodies [17, 8] from the music context to better assist with melody estimation. Such auxiliary information often contains useful hints for vocal melody extraction, but it is generally ignored in previous works. In this paper, we used two types of widely-accessible auxiliary information: *accompaniment* audio track and *lyric* timestamps, which leads to a context-assisted vocal melody transcription algorithm.

The accompaniment track contains the following useful music context:

1. Tonality and local chord context: Many semitone errors can be fixed by referencing to the tonality and chord information. For instance, if we detected a vocal note whose pitch is between F# and G in a C major chord, it is more likely to be a G note. Based on such preferences, semitone quantization error can be reduced.
2. Global tuning: The instruments, as well as the vocal parts, may not be fine-tuned [9], which can lead to systematic pitch errors if we refer to the

common standard where A4 = 440 Hz. To solve this problem, we can estimate the global tuning [13] from the accompaniment track and use it to better decode (de-trend) the vocal melody. It is worth noting that the global tuning algorithm by [13] works much better on the accompaniment track than on a pure vocal track because human pitch is more unstable compared to accompaniment instruments.

The above phenomena inspire us to introduce the accompaniment audio as an optional input to the system. Of course, we can adjust the vocal melody using the accompaniment in a rule-based way, but instead we let the neural network takes the accompaniment as an extra input and learn how it should constraint and adjust the melody in a data-driven way.

Lyrics are another common auxiliary resource for vocal melody audio. There are three typical formats of lyrics with different levels of timing accuracy, as listed in Table 1.

Lyric Format	**Common Sources**
Raw plain lyrics (no timestamps)	Plain-text lyric databases
Raw phrase-level timestamps	Music streaming services, Music Video (MV) subtitles
Raw word-level timestamps	Karaoke video subtitles

Table 1: Different formats of auxiliary lyric information and their common sources.

In this paper, we focus on the third kind of lyric format that is most informative. Raw word-level timestamps contain the human-annotated onset (and sometimes also the offset) of each word (or each syllable in some cases). We consider the case that the timestamps are word-level and only word onsets are available. Lyric onsets are highly related to note onsets. They are not entirely equivalent, but highly interrelated in the following ways:

1. One word may correspond to one or more consecutive notes.
2. Adjacent words may not share a single note. This means the onset of a word indicates the onset of one new note.

In practice, human-annotated timestamps are not guaranteed to be accurate. Under laboratory condition, onset annotation by tapping may cause a timing error within 60ms by trained humans [3]. We found in a small set of Karaoke video subtitles that the timing error of the annotated onsets varies from song to song, mostly within a standard deviation of 20ms to 60ms. Thus, our model treats the input lyric timestamps as an approximate reference.

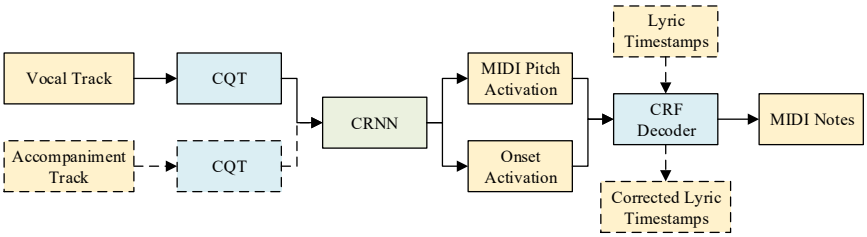


Fig. 2: System Overview. The accompaniment track and the lyric timestamps are optional inputs of the system.

2 Proposed Methods

2.1 System Overview

Our proposed system comprises of two components, the feature extractor, realized by a Convolutional Recurrent Neural Network (CRNN), and the decoder, realized by a Conditional Random Field (CRF). The system diagram is shown in Fig. 2.

The system first applies Constant-Q Transform (CQT) spectrograms for both the vocal track and the accompaniment track (if available) with a sampling rate of $f_{sr} = 22050$ and a hop length of $l_{hop} = 512$ samples. This CQT spectrogram is then processed by the CRNN to calculate the quantized pitch activation and word onset activation for each frame. After this, the activations and the raw lyric timestamps (if available) are treated as the observation of the CRF model. Finally, the CRF model jointly decodes the corrected lyric onsets and the note sequence. We use $N = 45$ distinct MIDI pitches from E2 (82.4 Hz) to C6 (1046.5 Hz) as the valid pitch range in the system.

2.2 Convolutional Recurrent Neural Network

The aim of the feature extractor is to obtain a representation of singing pitch and note onset that is irrelevant to different singing styles. To achieve this, we adopt a CRNN whose architecture is shown in Fig. 3.

The input of the model takes one or two channels (depending on whether the accompaniment spectrogram is available). The output of the model contains $(N+3)$ dimensions where the first $(N+1)$ dimensions are the frame-wise activations for N MIDI pitches and one non-voicing state (or the silence state). The last two dimensions are the activations for lyric onset detection. Both parts can be regarded as a classification task and are trained using the cross-entropy loss. The total loss of the model is computed as the sum of the losses from these two classifiers.

We use 8 convolutional layers with a small kernel size (3×3) with the rectified linear unit as the activation function. Batch normalization is performed between

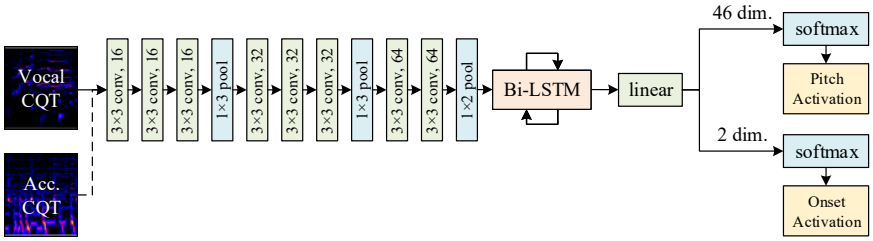


Fig. 3: Overview of the network architecture with $N = 45$ MIDI pitch classes. Accompaniment CQT is an optional input to the model.

convolutional layers. Three max-pooling layers with kernel size 1×3, 1×3 and 1×2 are used after the 3^{rd}, 6^{th} and 8^{th} convolutional layer. The output then goes through a Bi-directional Long Short-Term Memory (Bi-LSTM) layer with a hidden size of 256 for each direction.

When training the network, a 1000-frame segment is randomly selected from each song in one epoch. We train the model with Adam Optimizer [10] with an initial learning rate of 10^{-3}. If the performance on the validation set does not improve over 10 epochs, the learning rate is decreased from 10^{-4}, 10^{-5} to 10^{-6}. Training stops if the performance on the validation set does not improve over 50 epochs.

2.3 Conditional Random Fields

In this section, we describe the design of the decoding model. We first present some important insights of the decoder design:

1. The hidden state of each frame must contain the current voicing information (i.e., silence or voicing). Voicing states should contain the pitch information, while the silence state should also memorize its previous pitch in order to model the pitch transition probability between two consecutive notes as in Tony [12].
2. A note onset happens upon a state transition either from silence to voicing or between two different voicing states. Therefore, there is no need to represent the note onset state again with a distinct hidden variable.
3. Lyric timestamps can tell us the rough lyric onset positions. To exploit this information, we introduce the lyric position hidden states $I_{1...T}$, from which we can get the lyric onset positions. As discussed in section 1.2, we only allow a lyric onset happens when a note onset happens.
4. Notice that we allow approximate lyric timestamps as input. To deal with such inaccuracy, we do not force the i-th lyric onset happens exactly at the i-th timestamps. Instead, we penalize the distance between the actual lyric onset position and the given timestamp. This penalty helps the system to *correct* the raw lyric timestamps.

We now formally define our model. The CRF model is a joint decoding model of the target note sequence and the corrected lyric onsets. Fig. 4 illustrates the decoding process. We define the joint hidden variable $Z_t = (M_t, I_t, D_t)$ at frame t, where:

1. $M_t \in \{\pm m_1, \pm m_2, ..., \pm m_N\}$ is the note state. Each $m_i > 0$ denotes a valid MIDI pitch. A positive M_t means a voicing state with a MIDI pitch M_t and a negative M_t means a silence state with a previous MIDI pitch $-M_t$.
2. $I_t \in \{0, 1, 2, ..., K\}$ denotes the lyric position of that frame. $I_t = k$ means there are exactly k lyric onsets before frame $t+1$. We prohibit multiple lyric onsets on one frame, so either $I_t = I_{t-1} + 1$ or $I_t = I_{t-1}$ will hold for all $t > 1$. K is the total number of words.
3. $D_t \in \{0, 1, 2, ...\}$ denotes the distance (in frame) to the nearest previous MIDI onset.

The observation Y of the CRF model contains the pitch activation H and the lyric onset activation O given by CRNN, as well as the raw lyric timestamps L where L_i is the time of the i-th approximate lyric onset.

The CRF conditional probability is:

$$p(Z \mid Y) = \frac{1}{C}\psi_L(Z_1)\psi_R(Z_T)\prod_{t=2}^{T}\phi(Z_{t-1}, Z_t, Y) \tag{1}$$

Here, T is the total number of frames and C is the normalization coefficient. ψ_L and ψ_R are the endpoint constraint functions to force $I_1 = 0$ and $I_T = K$ so that the decoded sequence contains exactly K lyric onset frames. ϕ is the multiplication of six potential functions: four transition potential functions $\phi_L, \phi_T, \phi_B, \phi_D$ and two observation potential functions ϕ_O, ϕ_M. We will explain the form of all potential functions in the following sections.

Transition Potential Functions We use ϕ_L to model the lyric position transition:

$$\phi_L(I_{t-1}, I_t, L) = \begin{cases} \exp\left[-\left(\frac{t\delta - L_{I_t}}{\gamma_w}\right)^2\right] & \text{if } I_{t-1} + 1 = I_t \\ 1 & \text{if } I_{t-1} = I_t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Here, $\delta = l_{hop}/f_{sr}$ denotes the delta time between each frame. Intuitively, for each lyric onset ($I_{t-1} + 1 = I_t$), we penalize large difference between the actual lyric onset time t and the raw lyric timestamps L_{I_t}. Here, $_w = 0.065$ is the parameter that controls the degree of the penalty.

We use $\phi_T(M_{t-1}, M_t)$ to model the note state transition. For transitions to a silence state ($M_t < 0$), only pitch self-transition is allowed ($\phi_T(M_{t-1}, M_t) = 0$ if $|M_{t-1}| \neq |M_t|$) as silence states are designed to memorize their previous pitch. For transitions to a new voicing state ($M_t > 0$), ϕ_T uses a transition matrix

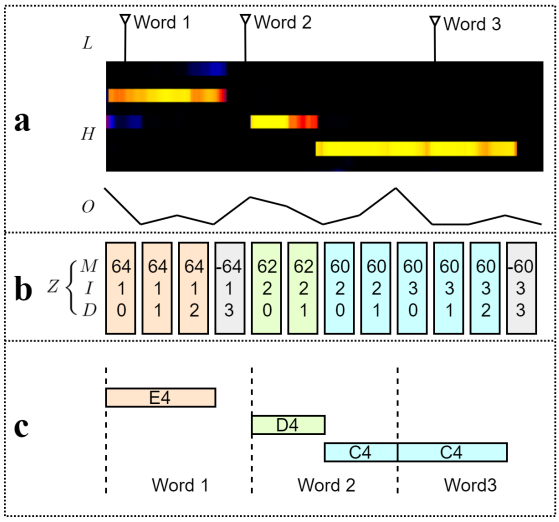


Fig. 4: An illustration of the decoding process: (a) the pitch activation H (shown by a spectrogram where lighter color denotes larger values) and the onset activation O are calculated as the observation of CRF, together with the approximate lyric timestamps L (if available); (b) the hidden variables (M, I, D) are decoded; (c) we construct the output note sequence and corrected lyric onset timestamps.

of successive note pairs. The transition matrix is trained from the MIDI note sequence in the training data.

We use ϕ_B to regularize the consistency of lyric onsets and MIDI note onsets. Specifically, $\phi_B(Z_{t-1}, Z_t) = 1$ if and only if any of the following conditions holds:

1. $M_t > 0 \wedge (I_{t-1} + 1 = I_t)$: The case of inter-word note transition.
2. $M_{t-1} > 0 \wedge M_t > 0 \wedge M_{t-1} \neq M_t \wedge I_{t-1} = I_t$: The case of inner-word note transition. Notice that no silence state is allowed during an inner-word transition, so we require $M_{t-1} > 0$ and $M_t > 0$.
3. $M_{t-1} = M_t \wedge I_{t-1} = I_t$: The case of self-transition.

Otherwise, $\phi_B(Z_{t-1}, Z_t) = 0$.

We use ϕ_D to restrict the minimal length d_{min} of a note to prevent too short notes accumulate near an observed onset peak. Specifically, $\phi_D(Z_{t-1}, Z_t) = 1$ if and only if any of the following conditions holds:

1. $(D_{t-1} + 1 \geq d_{min}) \wedge D_t = 0 \wedge \text{note\_onset}(t)$
2. $(D_t = D_{t-1} + 1) \wedge \neg\text{note\_onset}(t)$

Here, $\text{note\_onset}(t) = (I_t \neq I_{t-1} \wedge M_t \neq M_{t-1}) \wedge (M_t > 0)$. The first condition restricts the minimal previous note length $(D_{t-1} + 1)$ when a new note occurs. The second condition increases the D_t counter by 1 if no note onset occurs. If none of the conditions holds, $\phi_D\ (Z_{t-1}, Z_t) = 0$.

We introduce the potential function ϕ_D to serve as a hard restriction on decoded note lengths. Ideally, a relatively high self-transition probability in ϕ_T already regularizes the frequency of new note onsets. However, in practice, there are still some undesirable cases especially when the onset activation is over-smoothed and no lyric timestamp is provided. In this case, the model tends to decode excessive note onsets in a short period of time. While one way to solve the issue is by post-processing, we adopt another way by introducing the note length hidden states $D_{1...T}$ to record note lengths, and design the threshold condition in ϕ_D to suppress short notes when decoding.

Observation Potential Functions We use ϕ_O and ϕ_M to model the lyric onset observation and the pitch observation, respectively. ϕ_M directly outputs the pitch activation of current pitch state M_t. For ϕ_O, we use the following expression:

$$\phi_O(Z_{t-1}, Z_t, F_t) = \begin{cases} O_t, & I_{t-1} + 1 = I_t \\ e^{-\gamma_s}, & \text{otherwise} \end{cases} \tag{3}$$

Here, γ_s is a parameter that influences the estimated density of lyric onsets when lyric timestamps are not provided. The optimal $\gamma_s = 4.2$ is acquired by a grid search on 20 songs not included in the cross-validation dataset.

Model Inference We use the Viterbi algorithm to decode the optimal hidden state sequence for the model. In our implementation, equivalent states are combined into one, e.g., states with $D_t \geq d_{min}$ are equivalent to states with $D_t = d_{min}$ and other hidden parameters unchanged.

To reduce the state number brought by I_t, we calculate the raw lyric position I'_t with the approximate lyric timestamps and assume that I_t cannot be far away from I'_t, e.g., only states with $|I_t - I'_t| < k$ are considered (We set $k = 10$). The total time complexity is then reduced to $O(TN^2 d_{min} k)$, which is in general acceptable.

The model can be adapted to the situation when raw lyric timestamps are not provided by ignoring the ψ_L and ψ_R term. In this case, the original values of I_t become redundant and storing the parity of I_t is enough for decoding.

3 Experiments and Results

3.1 Dataset

We collected a dataset containing 1000 Chinese popular songs from the web. Each song in the dataset has a matched accompaniment track and a vocal track. The MIDI notes are annotated by human transcribers, and lyric onsets are aligned to them. The dataset is divided into five folds for cross-validation. We use three folds to train, one fold to validate and one fold to test. All songs in the training set are augmented by a pitch shifting from -4 semitones to 4 semitones and the MIDI pitch labels are changed accordingly.

3.2 Comparative Results

We first perform experiments on the system with and without accompaniment audio as an additional input. The first two measurements evaluate the precision, recall and F1 scores of frame-wise voicing/non-voicing detection and pitch classification, respectively. The note-level results (COnP, COnPOff) are acquired with the evaluation method from [16]. The onset tolerance is 50ms, and the offset tolerance is 30% of the note duration. The results are shown in Table 2.

		Tony [5]	**Cante [7]**	**CRNN**	**CRNN+A**
Voicing Segmentation	P	**0.9282**	0.8983	0.9044	0.9055
	R	0.8112	0.8445	0.9410	**0.9414**
	F1	0.8658	0.8706	0.9223	**0.9231**
Pitch	P	0.7246	0.7289	0.8193	**0.8228**
	R	0.6491	0.6853	0.8525	**0.8554**
	F1	0.6927	0.7064	0.8355	**0.8388**
COnP (Pitch, Onset)	P	0.4635	0.2942	0.7797	**0.8039**
	R	0.4492	0.3470	0.7646	**0.8117**
	F1	0.4563	0.3184	0.7721	**0.8078**
COnPOff (Pitch, Onset, Offset)	P	0.1842	0.1336	0.6489	**0.6801**
	R	0.1786	0.1574	0.6363	**0.6867**
	F1	0.1813	0.1445	0.6425	**0.6834**

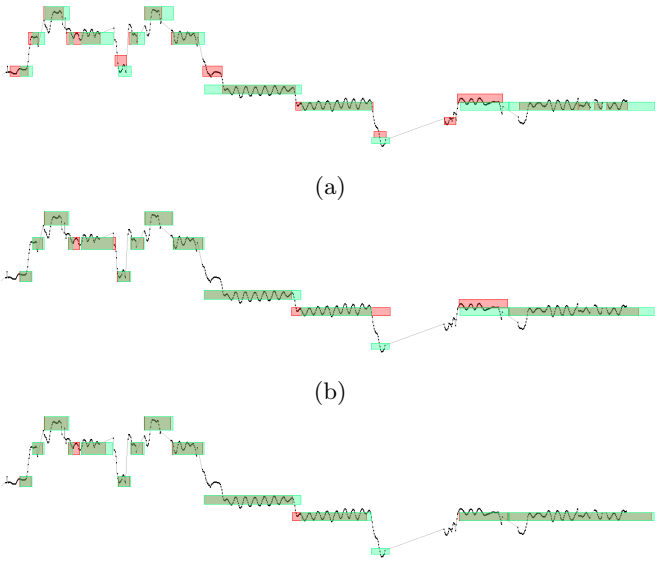
Table 2: A comparison between our method (CRNN) and former systems.

Here, CRNN denotes our proposed model without any auxiliary information, and CRNN+A denotes the model with the accompaniment track as additional inputs. We see that our method outperforms the baselines by 18% on the pitch F1 score and at least 77% on the note-level F1 score. Also, the presence of accompaniment audio further improves the performance as we expect. In Fig. 6, we show an example of how accompaniment audio will help improve the model accuracy.

3.3 Results with Approximate Lyric Timestamps

We further evaluate the system with the presence of raw lyric timestamps. To estimate the system performance against timing errors, a Gaussian noise with standard deviations σ_ϵ is manually added to the ground truth timestamps before decoding. The results are shown in Table 3 where L denotes the model with the raw lyric timestamps as additional inputs and the number in brackets denotes the σ_ϵ value we set.

We find the design of our CRF model useful for lyric onset correction by comparing it with another version that removes the lyric onset correction function and regards input lyric timestamps as true onsets. The results show that the performance of the model without lyric onset correction is greatly harmed



(a)

(b)

(c)

Fig. 6: A comparison of the decoded MIDI notes by (a) Tony [12], (b) CRNN without accompaniment information and (c) CRNN with accompaniment information to the same piece. With the presence of the accompaniment harmonic context, semitone errors between the output (*red blocks*) and the ground truth (*light green blocks*) can be further suppressed. The f_0 frequency is denoted by the black curve. Best viewed in color.

Models	**COnP F1 (w/ lyric onset correction)**	**COnP F1 (w/o lyric onset correction)**
CRNN	0.7721	/
CRNN+L (0ms)	0.8193	0.8600
CRNN+L (30ms)	0.8158	0.7699
CRNN+L (60ms)	0.7861	0.5629
CRNN+L (100ms)	0.6001	0.4754
CRNN+A	0.8078	/
CRNN+A+L (0ms)	0.8432	0.8663
CRNN+A+L (30ms)	0.8413	0.7786
CRNN+A+L (60ms)	0.8185	0.5773
CRNN+A+L (100ms)	0.6299	0.4906

Table 3: COnP results under different timing errors of the lyric timestamps.

when timing errors are introduced, yet the model with lyric onset correction maintained a high performance for $\sigma_\epsilon \leq 60$ms. In Fig. 7, we show an example of how lyric timestamps can help with decoding even if they are inaccurate.

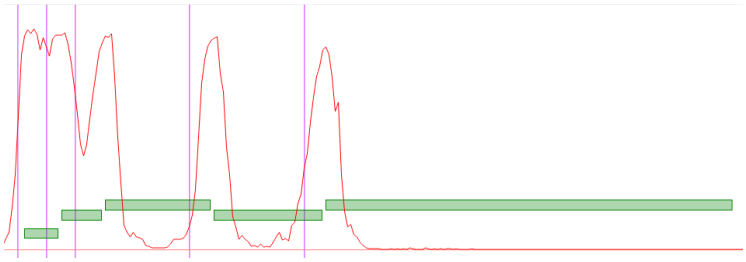


Fig. 7: An example of decoding with approximate lyric timestamps. The onset of decoded notes (*green blocks*) referenced the raw lyric timestamps (*purple vertical lines*), but some corrections are made according to the note onset activation (*red curve*).

In general, lyric timestamps still have a positive effect on the system performance when $\sigma_\epsilon \leq 60$ms. On the other hand, if the timestamps are barely accurate (e.g., $\sigma_\epsilon = 100$ms), it cannot be used as a good reference to the system.

4 Conclusion

In this paper, we present a novel system that greatly improves the accuracy for note-level monophonic vocal melody transcription with the Convolutional Recurrent Neural Network and the Conditional Random Fields. We also introduce two types of widely available auxiliary information, accompaniment audio tracks and approximate word-level lyric timestamps, to further improve the accuracy of the system. Future work will explore the adaption of the system to polyphonic music, as well as the utilization of other lyric formats, e.g., phrase-level timestamps, by combining the system with a lyric alignment model.

References

1. Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.
2. Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.
3. Roger B Dannenberg and Larry Wasserman. Estimating the error distribution of a tap sequence without ground truth. In *ISMIR*, pages 297–302, 2009.

4. Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968. IEEE, 2014.
5. Daniel PW Ellis and Graham E Poliner. Classification-based melody transcription. *Machine Learning*, 65(2-3):439–456, 2006.
6. Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by humming - musical information retrieval in an audio database. *Proceedings of the ACM International Multimedia Conference and Exhibition*, 10 1995.
7. Emilia Gómez, J Bonada, and Justin Salamon. Automatic transcription of flamenco singing from monophonic and polyphonic music recordings. In *III Interdisciplinary Conference on Flamenco Research (INFLA) and II International Workshop of Folk Music Analysis (FMA)*, 2012.
8. Shuhei Hosokawa and Toru Mitsui. *Karaoke around the world: Global technology, local singing*. Routledge, 2005.
9. Maksim Khadkevich and Maurizio Omologo. Phase-change based tuning for automatic chord recognition. In *Proceedings of DAFX*. Citeseer, 2009.
10. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
11. Sangeun Kum, Changheun Oh, and Juhan Nam. Melody extraction on vocal segments using multi-column deep neural networks. In *ISMIR*, pages 819–825, 2016.
12. Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. 2015.
13. Matthias Mauch and Simon Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1280–1289, 2010.
14. Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663. IEEE, 2014.
15. Brian McFee and Juan Pablo Bello. Structured training for large-vocabulary chord recognition. In *ISMIR*, pages 188–194, 2017.
16. Emilio Molina, Ana Maria Barbancho-Perez, Lorenzo J Tardón, Isabel Barbancho-Perez, et al. Evaluation framework for automatic singing transcription. 2014.
17. Vishweshwara Rao and Preeti Rao. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE transactions on audio, speech, and language processing*, 18(8):2145–2154, 2010.
18. Wei-Ho Tsai and Hsin-Chieh Lee. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1233–1243, 2012.
19. Luwei Yang, Akira Maezawa, Jordan BL Smith, and Elaine Chew. Probabilistic transcription of sung melody using a pitch dynamic model. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305. IEEE, 2017.

Extraction of Rhythmical Features with the Gabor Scattering Transform

Daniel Haider, Peter Balazs

Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, 1040 Vienna, Austria.

Abstract. In this paper we use the scattering transform to extract wide-scale information of musical pieces in terms of rhythmical features. This transform computes a layered structure, similar to a convolutional neural network (CNN) but with no learning involved. Applied to audio it is able to capture temporal dependencies beyond those possible for common time-frequency representations. This is already demonstrated by experiments for modulations of single tones. Here we provide a setup to include real world music signals, which extends the temporal range to the scale where rhythm and tempo live, allowing very intuitive explanations of how these scales are reached. In this way we also get an intuition of the mechanics inside a neural network when "listening" to rhythm.

Keywords: Gabor Transform, Scattering Transform, Rhythm, Tempo, Time-Frequency Representations, Convolutional Neural Networks

1 Introduction

A great goal in computer sciences these days seems to be making a computer to "listen" like humans do [1]. Among many machine learning approaches which have been aiming towards this, *Convolutional Neural Networks* (CNNs) perform particularly well, archieving impressive results in various audio-related learning tasks [2]. When such a network is tasked to learn from audio, it needs to identify and interpret patterns on several temporal scales: Pitch and timbre live within the scale of milliseconds, tempo and rhythmical structures are spread over periods of seconds and general progressions pass minutes and hours. To get a feeling of how this can be achieved, we briefly introduce CNNs and discuss the responsible mechanics. This motivates a special transform of audio signals, called *Scattering Transform*, which is computed by a cascade of time-frequency transforms setting up a layered network, similar to the structure of a CNN [3]. We emphasize on its ability to capture wide range dependencies of an audio signal and setup a framework to extract rhytmical structures of a musical piece in terms of its tempo. For this we use a scattering procedure based on a sampled Short-Time Fourier Transform (STFT), also called *Gabor Scattering* [4], which allows to illustrate particularly well, how coarser structures are captured and depicted in a very intuitive way. We may paste these insights to the CNN-case as a deterministic analogue.

2 Computers Listening - Convolutional Neural Networks

The idea of a neural network (NN) is to set up a function, that can theoretically approximate any other function via an optimization procedure. Such a network has a layered structure, each consisting of single neurons which filter the importance of the information arriving and a non-linear function, called *activation function* that controls the significance of the neuron to the network. Convolutional neural networks are a specialized form of NNs to deal with grid-like data, originally introduced in image processing. The idea is to convolve the input matrix with 2D filters that are much smaller than the input dimension, which can be interpreted as localization of certain properties of the data [5].
A feature extraction procedure usually yields the input data to a lower dimensional representation, thus, the dimensionality of the input data should be reduced in a meaningful way. This can be achieved by *pooling*. Pooling computes a "summary" of nearby elements, so it decreases the dimensionality and generates invariances to specific deformations and variations in the data. Furthermore, this expands the range of the filters in the subsequent layers since the filters are applied to the pooled "summary-elements", that are representative for a whole neighborhood of elements of the previous layer. Thus, the deeper the network gets, the wider dependencies are captured.
CNNs have led to an immense progress in image-related learning tasks. Clearly it makes totally sense to apply it also on the images obtained by time-frequency decompositions of audio signals [6]. Those decompositions already provide a representation of basic features of a signal, namely its frequency content, which is referred to as pitch, i.e. small-scale information. In the next section we will have a closer look at the scattering transform, a similar but non-learning construction based on cascades of time-frequency representations that expands the scale of the represented information.

3 Computers Listening Revisited - The Scattering Transform

Most of the time-frequency representations in practice are set up via a filterbank construction, i.e. a collection of filters, that decompose the signal into different frequency bins [7]. The filtering can be realized via the convolution operation; we write the discrete version as $y_k[m] := (x * w_k)[m] = \sum_n x[n] w_k[m-n]$, where x denotes the signal and w_k the filter corresponding to the k-th frequency bin. Usually, a modulus $|\,.\,|$ or a modulus squared $|\,.\,|^2$ is applied on the computed coefficients. Furthermore, some transformations also use dimensionality reduction in time, e.g. subsampling or scaling. This construction, i.e. filtering, taking the modulus and dimensionality reduction points out the fundamental link between filterbank decompositions and the principle structure of a CNN.
The Scattering Transform extends this link by computing a layered network structure, based on time-frequency decompositions with a modulus. In other words, time-frequency magnitude decompositions are applied on the single filter outputs of the previously decomposed filter outputs, see Figure 1.

Definition 1 (Scattering). *We compute a member of the ℓ-th layer L_ℓ of the scattering network by following a path $p = (p_1, \ldots, p_\ell)$ through its tree-like structure. Denoting the k-th filter in the ℓ-th layer as w_k^ℓ, then p denotes the indices of the used filters per layer, i.e. $w_{p_1}^1, \ldots, w_{p_\ell}^2$. So we can define,*

$$L_\ell[p] = ||\ldots|x * w_{p_1}^1| * \ldots| * w_{p_\ell}^\ell| \tag{1}$$

Originally, the transform was introduced by Mallat in [3] and was based on the wavelet transform. It came with a rigorous mathematical analysis, enabling it to show some translation invariance and stability w.r.t. time-deformations. Later the approach was generalized to semi-discrete frames, which include common filterbank constructions [8]. In [9] it was used as a feature extractor for audio and it turned out that it is able to represent temporal features beyond those possible for common time-frequency representations. Based on the pitch and filter structures contained in the first layer, the second layer reveals transient phenomena, such as note attacks and modulation of the amplitude and frequency. The examples there show that indeed, as we look at deeper layers of the scattering network, wider structures of the signal are represented. Applied on an amplitude modulated tone it is the envelope that appeared, i.e. the timbre of this particular sound and in a vibrato modulated tone it is the frequency of the vibrato pulse. In the next chapter we set up a framework to analyze rhythmical aspects of a musical signal and show that also these coarser temporal structures can be captured in the second layer of a scattering network.

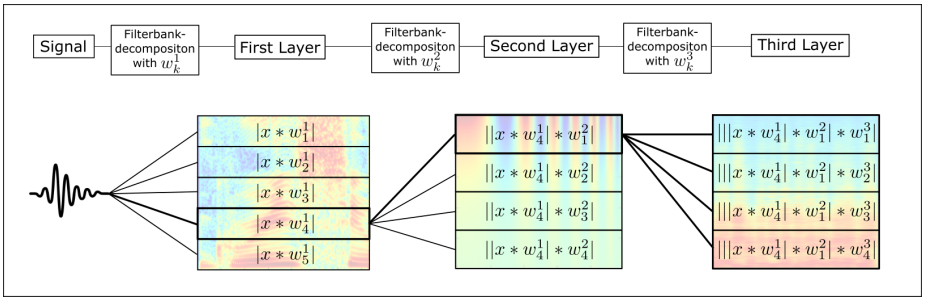


Fig. 1. The figure illustrates the scattering procedure by computing a third layer w.r.t. the path $[4, 1]$.

4 Rhythmical Feature Analysis

Rhythm refers to the timing of events within a musical piece and has different levels of periodicity. In the notation of western music, a hierachial metrical structure is used to distinguish between different time scales. The *Tatum* (temporal atom) is the smallest and is related to the shortest durational value encountered between two events within the musical piece. The *Tactus* (beat) is the perceptually most prominent and refers to the rate, most people would "tap" their feet to. Finally the widest, the *Bar* (measure) is related to the length of a rhythmical

pattern [10]. In a classical drumbeat in 4/4 with a bass drum on the 1 & 3, a snare on the 2 & 4 and an eighth note hihatpattern, Tatum would correspond to the hihat, Tactus to the snare and bass drum and the Bar to the whole pattern. The tempo of a musical piece is referred to the speed of the most prominent rhythm pattern, which is usually the Tactus. Tempo is measured in bpm (beats per minute) and reaches among different genres and styles of music from 30-300bpm. Embodying the tempo as a periodic pattern of events in time we could also assign a frequency to it; here in the range of 0.5-5Hz. As a frequency, this is clearly not audible, but indeed perceivable as a rhythmical pattern. The scattering transform is also capable of "perceiveing" a rhythmical pattern by depicting its (subsampled) frequency in the second layer. We explain this in particular, demonstrated on the most simple embodiment of tempo, a *metronome*.

4.1 Gabor-Scattering of a Metronome

We use a scattering transform based on a sampled Short-Time Fourier Transform (STFT) with a time-hop size parameter α, a window ϕ and a subsequent modulus operation applied. This is also known as *Gabor-Scattering* [4]. The following equation defines the sampled STFT applied on a signal x and shows, how it can be interpreted as filterbank decomposition w.r.t. filters w_k.

$$X_k[n] := \sum_m x[m] \underbrace{e^{-i\omega_k m}\phi[m-\alpha n]}_{=:w_k[\alpha n - m]} = (x * w_k)[\alpha n]. \tag{2}$$

We can model a metronome simply by an impulse train with periodicity T, i.e. a frequency at $1/T$,

$$e[n] = \sum_m \delta(n - mT),\ T \geq 0, \tag{3}$$

see Figure 2(a). We perform a STFT on e using a window, that is smaller than T to avoid overlapping. After applying a pointwise modulus, the first layer can be written as

$$L_1[k][n] = \sum_m |\phi(n\alpha - mT)| \tag{4}$$

for all w_k. This can be explained by viewing the convolution in Eqn. (2) as moving the filters w_k along single peaks of ones, which gives $\sum_m |w_k(n\alpha - mT)|$. The (pointwise!) modulus then removes the modulation part in w_k and only a sum of shifted windows remains. Eqn. (4) shows a smoothed and subsampled version of the metronome with a periodicity of T/α instead of T, i.e. a frequency of α/T, instead of $1/T$. If we choose α sufficiently large, then α/T can be depicted by a subsequent STFT in a more accessible frequency region. The second layer will thus show a constant frequency of α/T, see Figure 2. By sampling the time-frequency representation in wide time-steps we obtain downscaled signals in the filter bins, which are interpreted as having a higher frequency by subsequent filters. This enables to capture wider scales; the larger α is chosen, the wider the scale in focus. If we want to consider more complex musical signals, we may

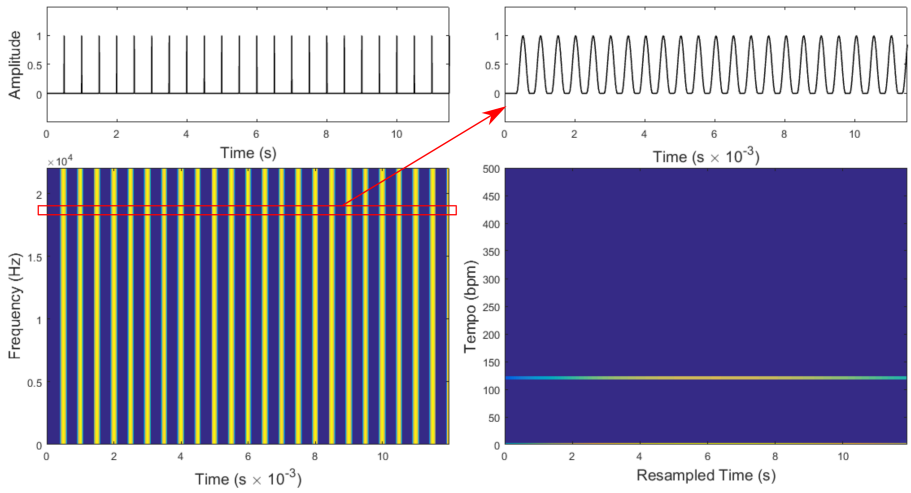


Fig. 2. Left: impulse train in the tempo of 120bpm, i.e. 2Hz and below its subsampled STFT, using a hop size of $\alpha_1 = 10^3$. Note that this is plotted here using a scaled timeaxis in seconds$\times 10^{-3}$! Right: the single filter outputs: the impulse train is smoothed by the window and scaled in time by the factor α_1, i.e. it is shorter and has a higher frequency, 2000Hz. Below is its STFT with $\alpha_2 = 1$ on a resampled timescale and a "tempo-scale" measured in bpm. It clearly depicts the tempo of the original signal train, 120bpm.

see them as consisting of tonal, transient and stochastic components. Truly, the transient parts of a signal indicate its rhytmical structure, e.g. by percussive elements like drums, note-onsets, etc. In that manner, the scattering transform will detect periodic patterns among the transient arrangements and extract the temporal information with respect to those. The measured level will depend on the most salient transient patterns in the signal. We set up three musical situations, where different levels of tempo are captured:

(a) A recording of a melody, played on a guitar in fingerpicking style by the first author. The melody consists of consecutive eighth notes, played rather monotonically, therefore the rhythmical structure, indicated by the onset-transients of the single notes yields a periodic pattern. Using a Tactus of 120bpm (referring to a quarter note beat), the Tatum of the melody is 240bpm. As the pattern is played rather monotonically, the most salient pattern is the one, induced by every note onset, i.e. 240bpm. Thus, this is the tempo, the second layer computes the highest values for, Figure 4(a) Right.

(b) A recording of chords, played on a guitar with hard palm-mute strokes by the first author. The chords are played in consecutive quarter notes in 150bpm (Tactus) with a strong accentuation on every fifth stroke. The rhythmical structure, indicated by the strokes has two levels now: on one side the Tactus

of the quarter notes in 150bpm and on the other side the Bar regarding to the accentuations in 30bpm. The strong accentuation makes both tempo levels accessible in the second layer of the transform, Figure 4(b) Right.

(c) An excerpt (2:45-3:20) of the song "Money" by Pink Floyd. It covers the transition from the saxophone solo (until 3:03) in around 120bpm into David Gilmour's guitar solo (from 3:07) in around 126bpm. The rhytmical structure is indicated by the drums, see Figure 3 below for the transcription. The hihat patterns usually defines the Tatum. In the saxophone solo, it consists of quarter notes over a 7/4 meter and since no finer scaled rhythm elements are present, the Tatum is 120bpm. In the guitar solo the pattern changes to triplets over a 4/4, i.e. the Tatum paces up to 378bpm. These values are clearly depicted in Figure 4(c) Right. The perceived tempo, the Tactus, is usually defined by the bassdrum/snare pattern. In the saxophone solo, Nick Mason plays the snare on the two, four and six with bassdrums inbetween, which yields an aperiodic pattern Bar-wise. Thus, the transform struggles to detect something meaningful for this level. In the guitar solo the pattern becomes periodic with 126bpm.

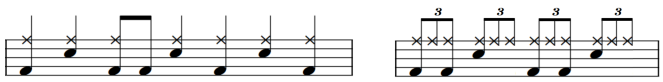


Fig. 3. Left: the drumpattern in the saxophone solo. Right: the drumpattern in the guitar solo. Crosses represent a closed hihat, F-notes the bass drum and C'-notes the snare.

We used time-hop sizes of 1000 samples for computing the first layers in (a),(b) and 2000 for (c). Different values amplify the access to different tempo levels, which makes it also possible to isolate certain levels by choosing the parameter appropiately. For the second layer, picking a single channel output can be problematic since we may catch one with a lot of tonal material, disturbing the detection of the transient pattern. As a preliminary solution we here computed the second layer using an average over channels corresponding to frequency regions where transients are most dominant. Of course, this also has an impact on the strenghts of certain tempo levels depending on the frequency distribution of the single transient sound. We used the channels corresponding to $1800-10300$Hz in (a),(b) and $1300-3500$Hz in (c).

The proposed procedure was not intended to be a tempo estimator as such, like [11] or [12]. The idea of this simple procedure is to extract temporal structures on several levels in their natural appearance. Other than common tempo estimators, salience is incorporated in the output as well, based on the energy of the transient sounds, which can be beneficial for further processing e.g. rhythm related tasks. Furthermore, it does not depend on external processing like beat tracking algorithms or a learning procedure, which makes it a simple, powerful tool.

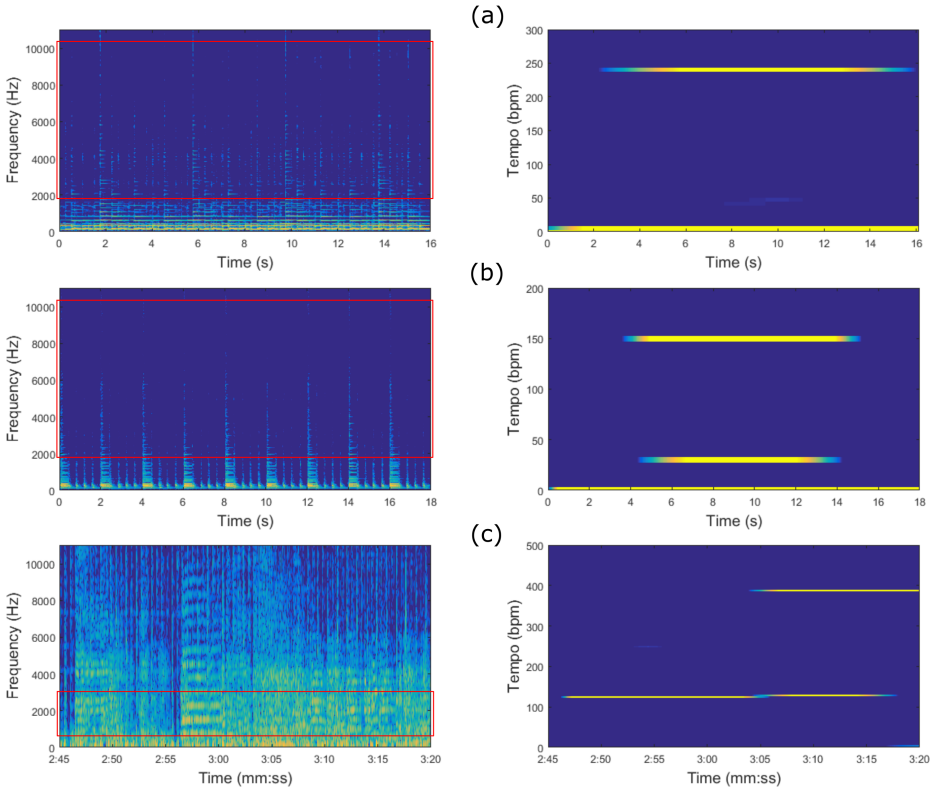


Fig. 4. Left: subsampled STFTs of the signals w.r.t. different (heuristically chosen) hop sizes α_1 and a modulus, plotted in seconds and Hz. The red areas indicate the channels used to compute the second layer. Right: common STFTs ($\alpha_2 = 1$) of the channel averages with modulus and averaging in time (second layer), plotted in seconds (referring to the original signal) and bpm.

5 Conclusion

We presented an rhythmic motivation of the scattering transform which extends the concept of time-frequency decompositions in a natural way. With a scattering transform based on a sampled STFT we ilucidated the mechanics behind the expansion of the captured temporal scales, explained on a simple impulse train. Examples in Figure 4 covered different musical situations with several tempo levels present, depicted in an intuitive way. As it is presented here, the preocedure is very preliminary but with potential to be expandable, fine-tuned and easily incorporated in other approaches. The insights into the network-like structure of the scattering transform can also give an intuition on the mechanics inside a CNN, when trying to learn the rhythmical structures of a musical piece.

Acknowledgements

The work on this paper was partially supported by the Austrian Science Fund (FWF) START-project FLAME (Frames and Linear Operators for Acoustical Modeling and Parameter Estimation; Y 551-N13). The authors thank Nicki Holighaus and Andrés Marafioti for fruitful discussions.

References

1. R. F. Lyon. Machine Hearing: An Emerging Field. *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131-139, 2010.
2. A. Marafioti, N. Holighaus, N. Perraudin and P. Majdak. Adversarial Generation of Time-Frequency Features with Application in Audio Synthesis. *arXiv*, `https://arxiv.org/abs/1902.04072`, 2019.
3. S. Mallat. Group Invariant Scattering. *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331-1398, 2012.
4. R. Bammer and M. Dörfler. Invariance and Stability of Gabor Scattering for Music Signals. *Proc. of Sampling Theory and Applications (Sampta)*, Tallin, Estonia. July 3-7, 2017.
5. I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning*, MIT Press, `http://www.deeplearningbook.org`, 2016.
6. J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann and X. Serra. End-To-End Learning for Music Audio Tagging at Scale. *Proc. of the 19th ISMIR Conference*, Paris, France. September 23-27, 2018.
7. P. Balazs, N. Holighaus, T. Necciari and D. T. Stoeva. Frame Theory for Signal Prcoessing in Psychoacoustics in: R. Balan, J. J. Benedetto, W. Czaja, M. Dellatorre, K. A. Okoudjou (eds.), *Excursions in Harmonic Analysis Vol. 5.* Basel (Springer), pp. 225-268, 2017.
8. T. Wiatowski and H. Bölcskei. Deep Neural Networks Based on Semi-Discrete Frames. *Proc. of IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, China. June 14-19, 2015
9. J. Andén and S. Mallat. Scattering Representation of Modulated Sounds. *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-12)*, York, UK. September 17-21, 2012.
10. A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*, Springer, 2007.
11. H. Schreiber and M. Müller. A Single-Step Approach to Musical Tempo Estimation Using a Convolutional Neural Network. *Proc. of the 19th ISMIR Conference*, Paris, France. September 23-27, 2018.
12. S. Dixon. Automatic Extraction of Tempo and Beat From Expressive Performances. *Journal of New Music Research*, vol. 30, no. 1, pp. 39-58, 2001.
13. P. Søndergaard, B. Torrésani and P. Balazs. The Linear Time Frequency Analysis Toolbox. *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 4, 1250032, 2012.

Kuroscillator: A Max-MSP Object for Sound Synthesis using Coupled-Oscillator Networks

Nolan Lem[1] and Yann Orlarey[2]

[1] Center for Computer Research in Music and Acoustics (CCRMA) Stanford University
[2] GRAME (Centre national de création musicale)
nlem@ccrma.stanford.edu

Abstract. This paper summarizes recent research using networks of coupled oscillators in real-time audio synthesis. We present two Max-MSP objects that synthesize the dynamics of these systems in real-time using both an additive and rhythmic synthesis model to generate complex timbre and rhythmic content. This type of self-organizing system presents many useful avenues of exploration in the field of sound synthesis and rhythmic generation. These objects allow users of Max-MSP to synchronize different ensembles of sinusoidal oscillators in real-time which can then be used as a vehicle for creative sound design, composition, and sound art.

Keywords: coupled oscillators, audio synthesis, Max MSP, generative music

1 Introduction

Coupled oscillator networks are a type of dynamical systems that describe a wide variety of interactive phenomena relevant to a number of research fields. Among those pertinent to the natural world, coupled oscillator systems have been used to account for firefly synchronization, synchronous chorusing in animal populations, and the cortical rhythms that comprise human neural networks [1]. In terms of musical beat and perception, Large and colleagues have incorporated coupled oscillators networks in their computational models to characterize how we become entrained to different types of rhythmic stimuli [2].

Previous research in coupled oscillator networks as a generative sonic device has looked at how their system behavior can be exploited in terms of their potential for creating musically relevant content [3] [4]. This includes exploring different strategies for rhythmic generation, audio synthesis, and as a control signal to approximate many compositional techniques found in contemporary music and computer music [5].

2 System Dynamics: Ensembles of Kuramoto Oscillators

In order to better understand the dynamics of these systems, a brief introduction to the Kuramoto model is presented. Kuramoto proposed a model of limit-cycle

oscillators that interact at the group level through their phase interactions [6]. Equation (1) shows the governing equation for such a system.

$$\dot{\phi_i} = \omega_i + \frac{K_i}{N} \sum_{j=1}^{N} sin(\phi_j - \phi_i) \tag{1}$$

where ϕ_i is the phase of the i_{th} oscillator and $\dot{\phi_i}$ is the derivative of phase with respect to time. ω_i is the intrinsic frequency of the oscillator, i, in a population of N oscillators. K_i is the coupling factor for each oscillator and the $sin(\phi_j - \phi_i)$ term is the phase response function that determines the interaction between each oscillator and the group. Typically, the range of intrinsic frequencies within the ensemble is taken from a Gaussian distribution, $g(\omega)$ at a center frequency, ω_c.

As K_i is increased, the oscillators with an ω_i closer to ω_c will begin to synchronize to the group by aligning their phases to other oscillators with similar frequencies. As more and more oscillators are recruited, synchrony emerges when $K_i > K_c$ where K_c is the point of critical coupling. Assuming a Gaussian distribution of intrinsic frequencies with a mean of ω_c, Kuramoto was able to show that as the number of oscillators goes to infinity, $K_c = \frac{2}{\pi g(\omega_c)}$.

Much more complicated types of synchrony occur when we let the intrinsic frequencies and coupling coefficients, $\omega_i(t)$ and $K_i(t)$, take on different values as a function of time. For example, Abrams and Strogatz were able to show the existence of "chimera states" that exhibit unusual dynamics where phase-locked oscillators coexist with asynchronous ones [7]. These system dynamics have perceptual implication in the audio synthesis routines described in the following section.

2.1 Programmatic Design

The main challenge in designing a real-time synthesis scheme using the aforementioned model is accounting for the phase interactions that must occur at each sampling interval. The phase response term in Equation (1) shows how the model grows exponentially as a function of N ($O(N^2)$) for each oscillator in the ensemble. To reduce the number of calculations per sampling interval, we use Kuramoto's application of mean-field coupling to the oscillators phases to derive the complex order parameters shown in Equation (2)

$$Re^{j\psi} = \frac{1}{N} \sum_{i=1}^{N} e^{j\psi_i} \tag{2}$$

where R and ψ are defined as the phase coherence and average phase respectively [6]. We can represent quadrature component of the complex phasor as a 90° phase shifted version of the in-phase part. Now the oscillators are no longer explicitly coupled to one another because their average phase governs their behavior. This is shown in Equation (3).

$$\dot{\phi}_{\mathrm{i}} = \omega_i + KR\sum_{j=1}^{N} sin(\psi - \phi_i) \tag{3}$$

From a computational standpoint, this has the benefit of reducing the number of calculations necessary to carry out the phase coupling adjustments and allows for a greater number of oscillators within each ensemble.

2.2 Faust Implementation

To create an interactive model, we utilized the functional programming language, Faust (Functional Audio Stream)[3], to implement the real-time signal processing. Faust is capable of being compiled into to a number of music programming related objects including Supercollider and Max MSP. Within Faust, the system is implemented by defining a series of "adjustable phasors" that can be modulated in terms of (instantaneous) phase and intrinsic frequency. By creating a feedback loop that calculates the average phase of the group of oscillators, the phase adjustments (shown in Equation (1) can be meaningfully applied to each term thereby allowing the oscillators phases to synchronize. This is shown in Fig 1 which shows the block diagram in Faust with four oscillators.

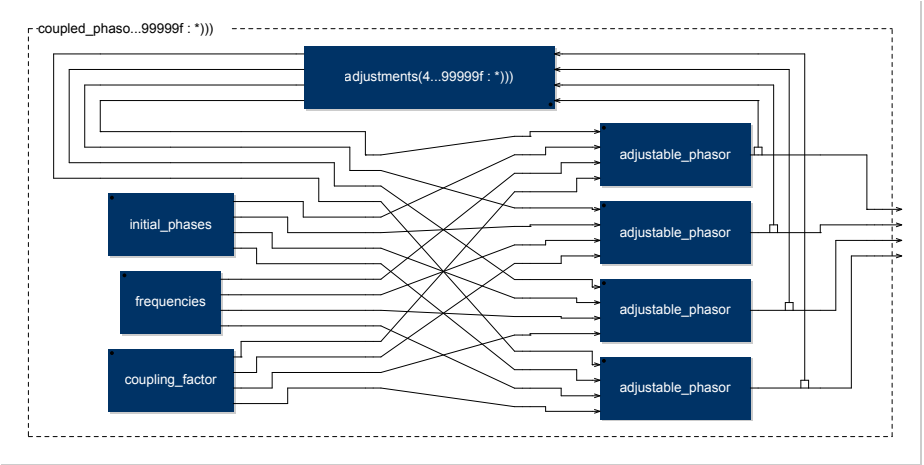


Fig. 1. Faust block diagram.

In the simple additive synthesis scheme, these phasors are applied as arguments into sine functions at the output to generate a bank of N phase-coupled sine waves. To use the oscillators to generate rhythms, we use the trajectory of

[3] `https://faust.grame.fr`

each phasor to trigger audio events upon each zero crossing ($\phi_i < \phi_{i-1}$) they encounter. We also must significantly decrease the distribution of ω_i to fall within a range normal for beat perception (0.25 Hz to 30 Hz). In order for the user to be able to interact with the model, each oscillator's coupling and intrinsic frequency are made variable.

3 Kuroscillator-rhythm and Kuroscillator-audio objects

The Kuroscillator max objects can be found in the directory listed below[4]. Figure 2 shows the *Kuroscillator-rhythm* and *Kuroscillator-audio* objects within the Max MSP programming environment.

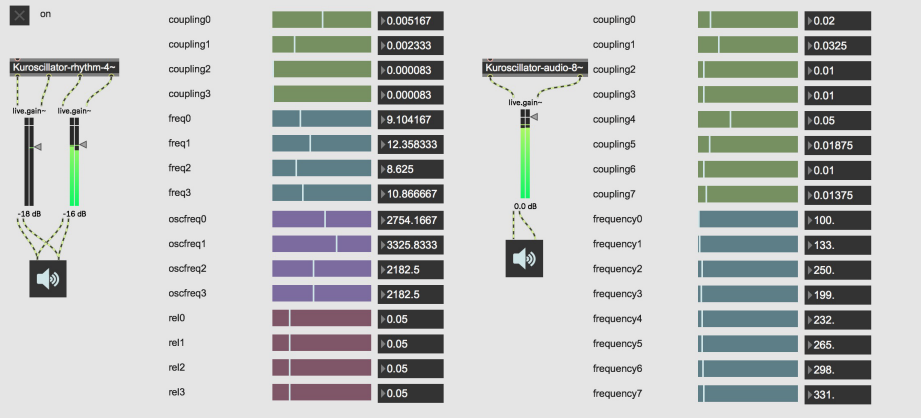


Fig. 2. *Kuroscillator-rhythm* object with 4 oscillators (left) and *Kuroscillator-audio* object with 8 oscillators (right).

Because the intrinsic frequencies of the oscillators in the Kuroscillator-audio object are themselves the audible frequencies at the output, their frequency range was limited to 0-5 kHz. In the Kuroscillator-rhythm object, the intrinsic frequencies determine the rhythm of the triggered sound and therefore they were limited a range just below the threshold of timbre formation, 0.25-30 Hz. In the Kuroscillator-rhythm object, the user can also modify the frequency of the audio oscillator's triggered note (33 Hz $< oscfreq <$5 kHz) and the length of the ASR envelope that gets applied to it. The Kuroscillator-rhythm object generates N individual outputs so they can be used as a control signal within Max MSP whereas the Kuroscillators-audio object is mixed down to two channels at the output. These constraints can be modified in the Faust source code that is located in the code directory.

[4] `https://bitbucket.org/no_lem/kuroscillators/src/master/`

The user defines which *type* of coupled oscillator object (*rhythm* or *audio*) and how many oscillators (N) in the ensemble using the convention "Kuroscillator-*type*-N" (where $N \leq 30$). Max MSP then generates the object and allows the user to interact with the coupling and intrinsic frequency by sending Max style messages to the Kuroscillator objects. These parameters are addressable: for the Kuroscillator-audio object, the user can send messages using the format "./Kuroscillator-type-N/coupling$i$ K_i" and ""./Kuroscillator-type-N/frequency$i$ $frequency$"" to modify the couplings and intrinsic frequencies respectively. For the Kuroscillator-rhythm object, the user can send additional Max messages of the form "./Kuroscillator-type-N/oscfreq$i$ *oscillator frequency*" and "./Kuroscillator-type-N/rel$i$ *sustain-time(sec)*". These parameters allow the user to explore different system states and bifurcations that are generated in real time. Assuming an intrinsic frequency distribution that is Gaussian (unimodal with respect to a center frequency), the oscillators can self-synchronize to the mean frequency of the distribution when they approach the critical coupling coefficient. Figure 3 shows the output of the *Kuroscillator-audio-30* object which shows a spectrum of a system of 10 oscillators synchronizing over a period of around 6 seconds.

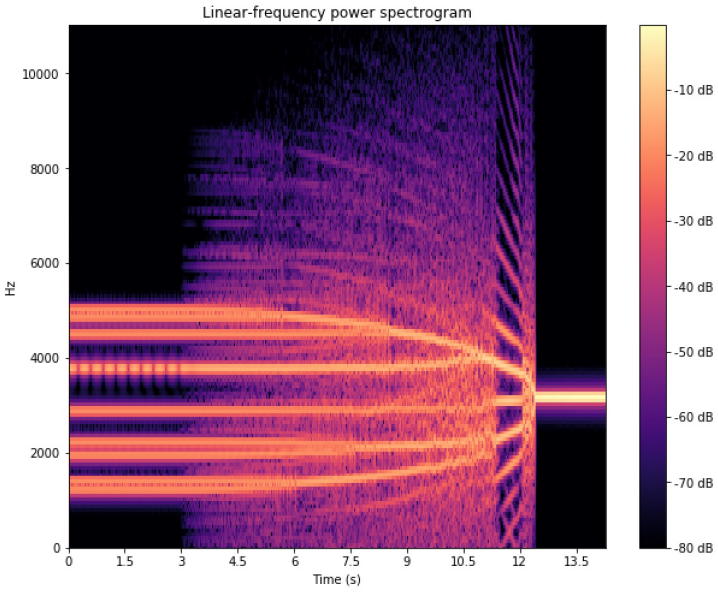


Fig. 3. Synchronizing frequencies of *Kuroscillator-additive* object with 30 oscillators.

Due to the flexibility of the programming environment, Max MSP is well suited for interaction with the model since the user can utilize the plethora of control objects in the Max tool kit. The examples directory hosted on the project site includes several Max example patches that show interesting system

states of synchrony by allowing control signals to modulate system parameters over time. These highlight the versatility of this model in producing sonic phenomena that approximates many pre-existing techniques already in the field of computer music namely (granular synthesis using microsounds) and contemporary composition (temporal canonization and phasing).

4 Conclusions

These coupled oscillator objects allow for real-time interaction with this particular self-organizing dynamical system. Because these systems contain a plethora of unusual output behaviors, Max MSP facilitates creative sonic exploration by allowing users to interact with system behavior using its output sound as a form of auditory feedback. Besides the sonic states characterized by "full synchrony" (in which all oscillators synchronize to a mean frequency), there exists a number of quasi-periodic states that emerge when oscillators take on different coupling coefficients and intrinsic frequency distributions [8]. In general, it is likely that the audio generated by these objects are perhaps best suited for types of music that are more oriented toward experimental music practices such as procedural, minimalist, and drone musical genres.

The Kuramoto model is just one particular type of coupled oscillator system that allows for self-synchronizing behavior. There are several other types of coupled systems that would be interesting candidates for synthesis: these include pulse-coupled oscillators (Mirollo-Strogatz oscillators), Van der Pol oscillators, and Stuart-Landau oscillators each one characterized by their own unusual dynamics [9] [3]. Future research on sonifying these dynamical systems would be enhanced by integrating existing psychoacoustic models that might allow us to better perceive the ways in which these complex systems behave.

References

1. Ravignani, A., Bowling, D., Fitch, W. T.: Chorusing, synchrony and the evolutionary functions of rhythm. In: Frontiers in Psychology. 5, 1–15 (2014)
2. Large, E., Herrera, J. A., Velasco, M. J.: Neural Networks for Beat Perception in Musical Rhythm. In: Frontiers in Systems Neuroscience, pp.1–14 (2015)
3. Lambert, A. a Stigmergic Model for Oscillator Synchronisation and Its Application in Music System. In: Proceedings of the International Computer Music Conference, pp. 247–252. Ljublana (2012)
4. Collins, N. Errant sound synthesis. In: Proceedings of International Computer Music. (2008)
5. Lem, N.: Sound in Multiples: Synchrony and Interaction Design using Coupled-Oscillator Networks. In: Proceedings International Conference Sound and Music Computing. Malaga, Spain (2019)
6. Kuramoto, Y.: Self-entrainment of a population of coupled non-linear oscillators. In: International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Physics, pp. 420–422. Springer, Berlin (1975)

7. Abrams, D., Strogatz, S.: Chimera States for Coupled Oscillators. In: Physical Review Letters, pp. 1–4. The American Physical Society, (2004)
8. Pikovsky, A., Rosenblum, M.: Dynamics of globally coupled oscillators: Progress and perspectives. In: Chaos. (2015)
9. Mirollo, R., Strogatz, S. Synchronization of Pulse-Coupled Biological Oscillators. In: SIAM Journal of Applied Mathematics, pp. 1645–1662.(1990)

Distinguishing Chinese Guqin and Western Baroque pieces based on statistical model analysis of melodies

Yusong Wu[1] and Shengchen Li[1]

Beijing University of Post and Telecommunications, Beijing 100876, P. R. China
wuyusong@bupt.edu.cn
shengchenli@bupt.edu.cn

Abstract. This paper proposes a method to determine different genres of melody according to the melodic interval of the melody with Western Barque and Chinese Guqin music used as an example. A melodic interval histogram and a Markov chain is proposed to differentiate Western Barque and Chinese Guqin music, where the similarity is measured with KullbackLeibler divergence. A significance test is done and the results shows that our method is capable of distinguishing between Western Baroque and Chinese Guqin pieces. This conclusion further supports that extracting melodic interval features could be a possible way to distinguish symbolic music melody from different genre.

Keywords: statistics, computational musicology, machine learning

1 Introduction

With the introduction of deep learning algorithms, automatic music composition has vastly developed in recent years. In the automatic music composition algorithms, a large proportion of the works are trained on a single music genre hence the generate music is the same genre with the music in training dataset. For instance, Bachbot [1] and DeepBach [2] compose polyphonic music in the style of Bachs Chorales; MidiNet [3] generates pop music melodies; Eck and Schmidhuber [4] proposed a system generating blues melody. Besides, there are models and systems that generates music in different genres. The WaveNet [5] is capable of generate music in different genre given conditional input, and the VRASH [6] system could generate melody in various genre with heuristic filtering.

Thus, it is interesting to differentiate the music genre of the melody for the validation of music genres generated by the auto music composition system. The purpose of this paper is to propose a method which could determine different music genres, where Bach Chorales music and Chinese Guqin music are used as an example.

In this paper, a computational method using interval histogram Markov chain is proposed to distinguish Western Baroque pieces and Chinese Guqin pieces. Melodies are analyzed in melodic interval sequence. A five-fold cross-validation

is applied and dataset is split into training set and test set. Firstly, a melodic interval histogram is drawn and a Markov chain is trained for the whole training set and on each music piece in test set. Then, for each music piece in test set, similarity using Kullback-Leibler divergence is measured between that music piece to each two genres. Last, to verify the effectiveness of our method, a paired t-test is used to verify the significant difference between the two similarity measured to two genres. The result of the paired t-test shows that both melodic interval histogram and Markov chain are capable of distinguishing Western Baroque pieces and Chinese Guqin pieces.

The remainder of the paper is organized as follows. Related works are first presented in section 2. Dataset and data representation used in this paper are described in section 3. Statistic models, algorithms are introduced in section 4. In section 5, experiment setups and experimental results are presented. Section 6 includes our explanations on the results. Conclusions are drawn and some future work is proposed in section 7.

2 Related works

In this paper, melody is represented as melodic interval. Melodic interval histogram and Markov chain are used to represent the melodic interval feature of the melody. Melodic interval refers to the distance between the two pitches of the two notes when played in sequence [7]. The melodic interval expresses the temporal sequence in terms of the distance between two adjacent pitches [8].

Representing symbolic music data in melodic interval often achieve good performance and widely used in music genre recognition (MGR) tasks. According to Correa [8], melodic interval representation offers more discriminative information than using absolute pitches or pitch contours. In [9], De Leon and Quereda compare the ability to separate two particular musical styles among a number of melodic, harmonic, and rhythmic statistical descriptors. The melodic intervals features in De Leon and Quereda's work achieve 100% test accuracy, De Leon and Quereda further conclude that melodic intervals features is one of the most discriminate features. In [10], Chai and Vercoe compared several representations of melodies in classifying European folk music. The representation used in Chai and Vercoe's work are absolute pitch representation, absolute pitch with duration representation, melodic interval representation, and contour based representation (contour based representation quantizes melodic interval representation into five levels, indicating conjunct motion, small ascending or descending in adjacent note, and large ascending or descending in adjacent note). By comparing the accuracy achieved by different data representation, Chai and Vercoe found that the interval representation outperforms the absolute pitch representation and the contour based representation. Besides, representing data as melodic interval also is invariant under pitch transpositions [8], thus giving the advantage of analyzing melodies regardless of the mode. Thus, we use melody interval for data representation in this paper, for not only the promising performance the

melodic interval got in previous works but also the merit of analyzing melody regardless of the pitch transpositions of the melody.

In addition to melodic interval representation, melodic interval histograms reflect the information about the order in note sequence and relative frequency of the intervals. Knopoff [11] uses melodic interval histogram to analyze the melodic activity in Bach's Fugue collection. Simsekli [12] uses a melodic interval histogram to classify music genre using bass lines on a 3-root 9-leaf label MIDI dataset. As much as accuracy of 100.00% for the root labels and 84.44% for the leaf labels are achieved, showing the effectiveness of melodic interval histogram extracting music feature.

Markov chains are also good at capturing temporal patterns in music. Verbeurgt [13] uses Markov chain to effectively extract patterns in music for composition.

3 Dataset and data representation

3.1 Dataset

The Western Baroque music and Chinese Guqin music are very unique music genre and have following characteristics that facilitate in distinguishing these two music genre: 1) The Western Baroque music and Chinese Guqin music have significant differences in music style which can be highly distinguishable by people even with little or no music education backgrounds. 2) In both Baroque music and Chinese Guqin music, most melodies do not contain overlap notes. That is, every note or chord ends before another chord or note starts. If multiple tones are played at the same time, they must start and end in the same time. This feature greatly facilitates the modeling and analysis for we can now form the melody as a sequence of notes and chords.

Thus, we choose Western Baroque music and Chinese Guqin music as an example to two different genres. Specifically, we use Bach Chorales dataset[1] and a self-collected Chinese Guqin dataset[2] as our dataset.

Bach Chorales dataset is a symbolic music dataset formatted in MuiscXML, containing 409 pieces, 7241 measures. Bach Chorales dataset is mostly in four parts harmony, containing four parts (Soprano, Alto, Tenor, Bass) in each score. In Bach Chorales dataset, some scores have extra accompaniment parts in addition to four chorales parts, such as violin, trumpet or timpani. The accompaniment parts often have few notes, with incomplete melodies and a large proportion of breaks. Because of the sparseness of the data in accompaniment parts, we only preserved four chorales parts, and all the accompaniment parts are ignored. Each of the four vocal parts is considered as a separate music piece.

The Chinese Guqin dataset is collected by ourselves, based on several books of score collection published by Chinese Guqin professionals. The scores are mainly

[1] Bach Chorales dataset: `https://github.com/cuthbertLab/music21/tree/master/music21/corpus/bach`

[2] Chinese Guqin dataset: `https://github.com/lukewys/Guqin-Dataset`

written in numbered notation, along with the Chinese Guqin notations[3]. A self-designed transcript system is used for typing the Chinese numbered notation into text format and convert the text format into MusicXML files. It is worth noting that, although the Chinese Guqin scores contain other notation of the music such as ornaments, fingerings, and expression notation, such information are ignored. Only the pitch and duration of melodies in Chinese Guqin scores is obtained. For scores containing multiple section, we regard each section as different music pieces. The Chinese Guqin dataset contains 247 pieces, 6107 measures in total.

3.2 Data representation

Although both melodic pattern and rhythmic pattern are different between Western classical music and Chinese classical music, the performance of Chinese Guqin music is very expressive, such that the duration of the note would varies in different Guqin player and play style. Thus, we focus primarily on extracting characteristics of tonal feature in Western classical music and Chinese classical music, regardless of the rhythm pattern in melody.

Western music system is based on heptatonic scale including seven levels per octave in scale, while the Chinese music system is mostly in pentatonic scale including five levels per octave, derived from the cycle-of-fifths theory. There are different mode variants (such as major or minor) in both Western music and Chinese music. Thus, by using a melodic interval representation which is invariant under transpositions, we could represent the melody regardless of their mode.

In this paper, we represented melodies in melodic interval sequence. We define the melodic interval as the absolute value of pitch difference in semitone. The melodic intervals are denoted in 13 classes, indicating semitones of intervals range in [0,12]. Intervals that are not multiple of 12 would be modulo by 12, following the method used in multiple works [10] [11]. However, unlike previous works using positive and negative intervals to denote upward and downward of the melodies, we consider only the absolute value of pitch difference, for reducing the dimension of the distribution subsequently extracted. Also, for intervals which are multiple of 12 and not zero, they are indicated as 12, as to capture the pure octave leap. Only melodic interval of 0 semitone would be recorded as 0, to capture the conjunct motion, i.e. repeated notes, in melody progression.

Given two adjacent pitch (p_t, p_{t+1}) in melody, where p_t represents pitch at time t, we define the melodic interval i_t at time t and the function of computing the interval $\mathrm{INT}(p_t, p_{t+1})$ as follows:

$$\mathrm{INT}(p_1, p_2) = i_t = \begin{cases} |p_{t+1} - p_t| \bmod 12 & \textbf{if} |p_{t+1} - p_t| \nmid 12 \\ 0 & \textbf{if} |p_{t+1} - p_t| = 0 \\ 12 & \textbf{otherwise} \end{cases} \quad (1)$$

[3] The introduction of Chinese numbered notation Chinese Guqin notations can be found at `https://en.wikipedia.org/wiki/Numbered_musical_notation` and `https://en.wikipedia.org/wiki/Guqin_notation`

As we mentioned above, all the melody in our dataset can be written as sequence of notes and chords. Since we only care about the melodic progression, i.e. pitch progression, we can represent notes and chords as pitch set, regardless of the duration of notes and chords. A note, which includes only one pitch, would corresponds to a pitch set containing only one pitch; a chord, which corresponds to multiple pitches, would corresponds to a pitch set containing multiple pitches. We denote the pitch set at time t as P_t. Thus, we can form the sequence of notes and chords into sequence of pitch sets $\{P_1, P_2, P_3 \dots P_m\}$.

Given a pair of pitch set containing two adjacent pitch (P_t, P_{t+1}), the interval set for such pitch pair I_t is defined as the set of all possible combination of progression from one pitch in P_t to next pitch P_{t+1}, considering all the transitions in note pitch contains indispensable information of melodic progression.

In other words, given a pitch set (P_t, P_{t+1}), the interval set of the pitch set would be:

$$I_t = \{\text{INT}(p_i, p_j)\} \quad \forall p_i \in P_t \ , \ \forall p_j \in P_{t+1} \tag{2}$$

One example of our data representation are shown in Fig. 1. The pitch set representation of the example, in MIDI absolute pitch, is $\{[60], [57], [57, 45], [64], [62], [60], [62], [64, 52]\}$, each square bracket represents a pitch set and each number in square bracket represents the pitch in the pitch set. The melodic interval is calculated as $\{[\text{INT}(60, 57)], [\text{INT}(57, 57), \text{INT}(57, 45)], [\text{INT}(57, 64), \text{INT}(45, 67)], [\text{INT}(64, 62)], [\text{INT}(62, 60)], [\text{INT}(60, 62)], [\text{INT}(62, 64), \text{INT}(62, 52)]\}$, which in result is $\{[3], [0, 12], [7, 7], [2], [2], [2], [2, 10]\}$, each square bracket represents a interval set and each number in square bracket represents the interval in the interval set. The value of the number denotes the interval value in semitone.

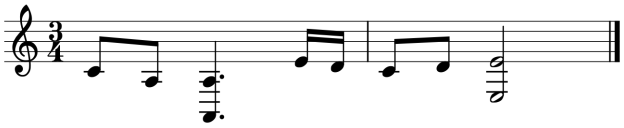


Fig. 1: One music segment example for data representation. The melodic intervals are computed using all possible combination of the difference in pitch for adjacent note. The melodic interval sequence for this music segment example would be: $\{[3], [0, 12], [7, 7], [2], [2], [2], [2, 10]\}$. Each square bracket represents a interval set and each number in square bracket represents the interval in the interval set. The value of the number denotes the interval value in semitone.

4 Method

4.1 Melodic interval histogram

We extract the melodic interval histogram as one feature. The distribution of melodic intervals are extracted in melodic interval histogram.s By comparing the probability in different interval, we can examine how music in different genres are different in the use of melodic interval and melodic progression.

For each genre, the melodic intervals are counted on all the music pieces in that genre. The frequency of melodic interval is then normalized, formed as a histogram. For genre G, the melodic interval histogram IH is calculated as follows:

$$\mathrm{IH}^G\{x = i\} = \frac{\mathrm{count}(x = i)}{N_I} \tag{3}$$

where N_I is the total number of melodic intervals counted and $\mathrm{count}(x = i)$ denotes the number counted for interval i, $i = 0, 1, 2 \ldots 12$.

4.2 Markov chain

Although the melodic interval histogram contains statistic features of a music genre, the information in melodic interval histogram is limited in static aspect. In other words, the melodic interval histogram only reveals the pattern in how melodic intervals are distributed, but not include how the melodic intervals are progressed. To extract temporal pattern of the intervallic progression, we trained a Markov chain using the melody on each genre of music. By analyzing the transition matrix in the Markov chain, we could extract pattern in melodic interval transition, i.e., the pattern of the melodic interval progression.

A Markov chain is a stochastic model describing the probability of the next state depending only on the current state. The probability of transferring from state i to state j is called transition probability p_{ij}, namely:

$$p_{ij} = \Pr\left(X_1 = j | X_0 = i\right) \tag{4}$$

We train the Markov model by counting the transition number and then normalize the transition count to transition probabilities. After we trained a Markov model, a parameter matrix is obtained containing all the transition probability. The parameter matrix is called the transition matrix of the Markov model. The transition matrix $P = (p_{ij})$.

In this paper, for each genre, a Markov chain is trained using all the music pieces in that genre. The state of Markov chain is the melodic interval, consisting 13 states representing melodic intervals from 0 to 12, in semitone. In other words, the Markov chain is trained on melodic interval sequence. In our melodic interval sequence, multiple interval occurs in one timestep (as interval set). To resolve this problem, we count all the possible combination of melodic interval transition, similar to how we define interval set.

After the Markov chain is trained, the transition matrix $\mathbf{M}$ is obtained. We use $\mathbf{M}$ instead of P to denote transition matrix for distinguishing from the pitch set in previous section. As a total 13 melodic intervals are used in melodic interval representation in this paper, the dimension of transition matrix is $\mathbf{M}$ is 13×13.

4.3 Similarity evaluation using KullbackLeibler divergence

With the two probabilistic model being built on the melodies in two genre respectively, for each genre, two distributions are obtained, namely melodic histogram and transition matrix in Markov chain. In other words, IH^{Bach}, $\mathbf{M}^{Bach}$, IH^{Guqin}, and $\mathbf{M}^{Guqin}$ are obtained. The superscript on melodic interval histogram IH and transition matrix $\mathbf{M}$ denotes the source on which the distribution is calculated. In this way, the melodic interval pattern of given music genres can be represented in melodic histogram and transition matrix. Given a new piece of music melody S, we can extract the same two distributions on them, namely interval distribution IH^S and interval transition matrix for that music piece $\mathbf{M}^S$, and measure how similar the distributions of the given music piece are from the distributions extracted from Western Baroque pieces and Chinese Guqin pieces.

KullbackLeibler divergence is algorithm used to measure the similarity between two distribution. Using KullbackLeibler divergence to measure the similarity between the distribution of the feature can serve as a intuitively and explainable method to measure the similarity.

Let $P(x)$ and $Q(x)$ be two probability distributions of a discrete random variable x. The KullbackLeibler divergence between P and Q is defined as follow:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{5}$$

The KullbackLeibler divergence is always non-negative. The more similar two distribution are, the less KullbackLeibler divergence value is computed. In other words, the lesser the KullbackLeibler divergence value, the higher the similarity between the two distribution. If two distributions are identical, the KullbackLeibler divergence computed between the two identical distributions would be 0. The KullbackLeibler divergence is not symmetric. The "$\|$" in the notation of KullbackLeibler divergence denotes this asymmetric, as $D_{\mathrm{KL}}(P\|Q)$ measures the amount of information lost when Q is used to approximate P.

In this paper, we define the similarity of sample S to certain genre G as the KullbackLeibler divergence from the distribution of feature extracted from sample D^S and the same kind of distribution of feature extracted from genre D^G. Here, distribution D is one of melodic interval histogram or transition matrix, specifically, $D \in \mathrm{IH}, \mathbf{M}$. The similarity is computed as follows:

$$D_{\mathrm{KL}}(D^G\|D^S) = \sum_{x\in\mathcal{X}} D^G(x)\log\left(\frac{D^G(x)}{D^S(x)}\right) \tag{6}$$

In practice, before computing the KullbackLeibler divergence of two distributions, a small number of $1e-5$ is added to each entry in both two distribution.

By smoothing the distribution, we can avoid subsequent calculation of Kullback-Leibler divergence on zero entries.

5 Experiments and Results

5.1 Cross-validation

A five-fold cross-validation is used to evaluate our algorithms. All the pieces are randomly shuffled and split into five subsets with approximate same length of total measures. Training and testing are practiced five times. For each time, we use the one subset as test set and the other four subsets as training set. For each genre, we build a melodic interval histogram and train a Markov chain using the music pieces in the training set. That is, we calculate $\mathrm{IH}^{Bach}, \mathbf{M}^{Bach}, \mathrm{IH}^{Guqin}$, $\mathbf{M}^{Guqin}$ on entire training set of each genre. For each music piece in test set, the interval histogram and Markov chain are built and IH^{S}, $\mathbf{M}^{S}$ are obtained. Then, KullbackLeibler divergence of $D_{\mathrm{KL}}(\mathrm{IH}^{Bach} \| \mathrm{IH}^{S})$, $D_{\mathrm{KL}}(\mathrm{IH}^{Guqin} \| \mathrm{IH}^{S})$, $D_{\mathrm{KL}}(\mathbf{M}^{Bach} \| \mathbf{M}^{S})$, $D_{\mathrm{KL}}(\mathbf{M}^{Bach} \| \mathbf{M}^{S})$ are computed.

5.2 Significance test using paired t-test

If our method is able to extract distinguishable feature, given a sample from test set S^G of genre G, the similarity of the sample to the genre G should be significantly higher than the similarity to the other genre G', $G' \ni G$. In our case, the KullbackLeibler divergence value is inversely proportional to similarity. Thus, given a test sample S^G from genre G, and distribution D^S extracted from genre S, we should have:

$$D_{\mathrm{KL}}(D^G \| D^S) < D_{\mathrm{KL}}(D^{G' \ni G} \| D^S) \tag{7}$$

A paired t-test is applied to verify the significant differences between KL divergence on two genres, and a one-tail p-value is calculated to examine the significance of the hypothesis above. We choose the significant level of $p = 0.01$. That is, if the p-value calculated from the paired t-test is lower than 0.01, we can say for a given sample $S \in G$, there is statistically significant that $D_{\mathrm{KL}}(D^G \| D^S)$ is larger than $D_{\mathrm{KL}}(D^{G' \ni G} \| D^S)$.

The paired t-test is applied as follows. A set of paired data with n pairs X_i and Y_i are given. In our case, a single pair X_i and Y_i are the similarity measured from one distribution extracted from a music piece to each two distributions extracted from two music genre, namely $D_{\mathrm{KL}}(D^G \| D^S)$ and $D_{\mathrm{KL}}(D^{G' \ni G} \| D^S)$. First, the mean difference $\bar{d} = Y_i - X_i$ of the pairs is computed. Next, the standard deviation of the differences s_d is calculated, and using s_d, standard error of the mean difference is calculated as $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$. Then, the t-statistic t is calculated as $t = \frac{\bar{d}}{SE(d)}$. The t-statistic t follows a t-distribution with degrees of freedom $df = n - 1$, and comparing t to the t_{n-1} distribution, the p-value p is obtained. The p-value p , is the probability that the test statistic will take a value at least as extreme as the observed value, assuming that the null hypothesis is true.

5.3 Results

The melodic interval histogram of Bach Chorales collection and Chinese Guqin collection on the whole dataset are shown in Fig. 2. From the melodic interval histogram in Fig. 2, we can see the major difference between Guqin collection and Bach Chorales collection is the use of the minor second and minor third. The Bach Chorales collection is abundant in the minor second while the Chinese Guqin music is more preferable to the minor third. Also, from the melodic interval histogram, we can observe that Bach Chorales collection is relatively extensive in the intervals of 0, 1 and 2 semitones, while having some proportion of intervals of 3, 4, 5, 7 and 12 semitones. In Chinese Guqin music, the proportion of the interval of 12 semitones is relatively high.

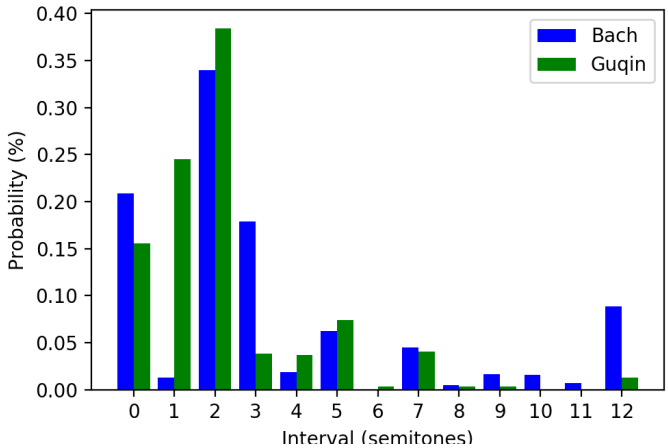
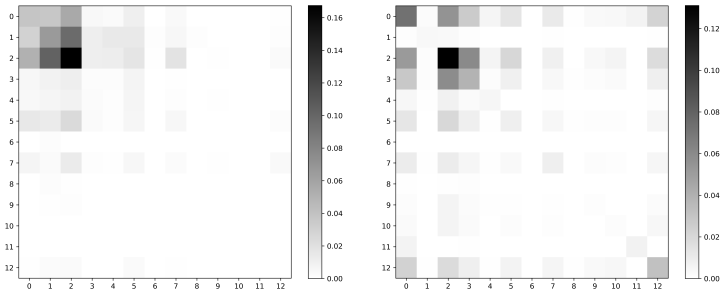


Fig. 2: The melodic interval histogram of Bach Chorales collection (left) and Chinese Guqin collection (right). Each bar represents an interval value and the height of the bar is the probability of the corresponding interval.

The transition matrix in Markov model trained on the whole dataset of Chinese Guqin and Western Baroque pieces are shown in Fig. 3. From the transition matrix shown in Fig. 3, we can observe that the melodic interval transition distribution in Bach collection concentrates on the upper left of the matrix, while the interval transition distribution in Guqin collection is more separated. This reflect that the Bach collection have less large intervals, while the melodic progression in Chinese Guqin collection involves more big leaps in melodic intervals.

The paired t-test result for KullbackLeibler divergence result is presented in Table. 1a and Table. 1b. The mean-difference($\bar{d}$), degrees-of-freedom(df), statistic(t) and the one-tail p-value(p) of each time of cross-validation is presented in two tables. The number followed the genre represents the number of



(a) Bach (b) Chinese Guqin

Fig. 3: The transition matrix of Bach Chorales collection (a) and Chinese Guqin collection (b). The number in the horizontal axis and vertical axis represent the interval value in semitone. The darkness of the element in the matrix denotes the probability of the transition.

the five-fold cross-validation. The results show that a p-value less than 0.01 is obtained in every significance test. This suggest that for a given sample in test set, $D_{\mathrm{KL}}(ID^{G}\|ID^{S})$ is significantly lower than $D_{\mathrm{KL}}(ID^{G'\ni G}\|ID^{S})$, and $D_{\mathrm{KL}}(\mathbf{M}^{G}\|\mathbf{M}^{S})$ is significantly lower than $D_{\mathrm{KL}}(\mathbf{M}^{G'\ni G}\|\mathbf{M}^{S})$. The results also show that transition matrix is better in distinguishing melodies, as the table of transition matrix results in larger mean-difference and lower p-value.

6 Discussion

Our result on melodic interval histogram of Bach Chorales music is similar to Knopoff and Hutchinson [11], who found that the conjunct movement is preponderance in Bach's fugue pieces, and movements by a major or minor second constitute about 70% of all pitch movements. Also, from results on melodic interval histogram, the lack of minor second in Chinese Guqin collection is mainly due to the musical scale in Chinese music pentatonic scale system dose not contain minor third. The abundant of interval of 12 semitones shows that Chinese Guqin music has a unique preference of using ocatave leap in melodic progression.

The sparseness differences of the transition matrix between Western Bach Chorales music and Chinese Guqin music confirm the intuitive fact that Bach collection often have more consecutive melody line, whereas the Chinese Guqin collection are jumpy in melodic progression. This reflects the difference in mode system between Chinese classical music and Western classical music. Also, the jumpy in Chinese Guqin melody is partly because it is commonly used in Guqin composing to raise or lower the notes by octaves in melody [14]. Interestingly, both two genres have maximum on the transition from the interval of second to the interval of the second.

(a) Melodic interval distribution

	$\bar{d}$	df	t	p
Guqin 1	-1.72	47	-13.52	2.03e-18
Guqin 2	-1.88	48	-20.45	6.03e-26
Guqin 3	-1.81	48	-16.15	1.24e-21
Guqin 4	-1.67	49	-14.40	7.75e-20
Guqin 5	-1.80	56	-16.40	1.20e-23
Bach 1	0.66	79	28.02	2.14e-43
Bach 2	0.64	85	44.52	2.75e-61
Bach 3	0.66	77	29.64	1.82e-44
Bach 4	0.68	80	36.74	1.85e-52
Bach 5	0.64	81	37.29	2.28e-53

(b) **Transition matrix**

	$\bar{d}$	df	t	p
Guqin 1	-1.83	47	-13.11	6.51e-18
Guqin 2	-2.22	48	-26.54	6.13e-31
Guqin 3	-2.10	48	-17.82	2.12e-23
Guqin 4	-2.06	49	-19.05	6.68e-25
Guqin 5	-2.23	56	-21.25	4.40e-29
Bach 1	1.98	79	51.46	3.99e-63
Bach 2	2.06	85	54.63	1.38e-68
Bach 3	2.03	77	48.72	3.20e-60
Bach 4	2.06	80	69.26	8.91e-74
Bach 5	1.95	81	51.95	1.45e-64

Table 1: The paired t-test result for KullbackLeibler results of melodic interval distribution (a) and transition matrix (b). The mean-difference ($\bar{d}$), degrees-of-freedom (df), statistic (t) and the one-tail p-value (p) are presented. The number followed the genre represents the number of the five-fold cross-validation. The transition matrix is better in distinguishing melodies, as the table of transition matrix results in larger mean-difference and lower p-value.

From both Table. 1a and Table. 1b, we can see a p-value much lower than 0.01 in each paired t-test. The low p-value supports that it is statistically significant that, given a music piece, the similarity measured to one genre is different from the similarity measured to another genre. This statistical conclusion suggests that both two distributions extracted from two genre are significantly different from each other. Thus, both melodic interval histogram and Markov chain are capable of capturing unique characteristics in melodies in Chinese Guqin and Western Baroque pieces and distinguish between the two music genres. The mean difference in paired t-test of transition matrix is larger than the mean difference in paired t-test of melodic interval histogram, which suggests that Markov chain is more distinguishable or separable than melodic interval histogram. Thus, Markov chain is better than interval histogram in differentiating Western Baroque pieces and Chinese Guqin pieces.

7 Conclusion

By building melodic interval histogram and Markov chain, features of melodic intervals are successfully extracted from melodies in Chinese Guqin and Western Baroque pieces. KullbackLeibler divergence is used to compute the similarity of given music piece to music genre. A significance test is done and the results show that our method are capable of distinguishing Western Baroque and Chinese Guqin pieces. The Markov chain performs better in distinguishing Western Baroque and Chinese Guqin pieces. The success of our model further suggest that melodic interval feature can serve as an effective way to extract characteristics of symbolic music melody.

References

1. Feynman Liang. Bachbot: Automatic composition in the style of bach chorales. *University of Cambridge*, 8:19–48, 2016.
2. Gaëtan Hadjeres, François Pachet, and Frank Nielsen. Deepbach: a steerable model for bach chorales generation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1362–1371. JMLR. org, 2017.
3. Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *Proceedings of the 18th International Society for Music Information Retrieval Conference ISMIR 2017*, 2017.
4. Douglas Eck and Juergen Schmidhuber. A first look at music composition using lstm recurrent neural networks. Technical report, IDSIA USI-SUPSI Instituto Dalle Molle, 2002.
5. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
6. Alexey Tikhonov and Ivan P Yamshchikov. Music generation with variational recurrent autoencoder supported by history. *arXiv preprint arXiv:1705.05458*, 2017.
7. Theta Music Trainer. Melodic intervals. `https://trainer.thetamusic.com/en/content/melodic-intervals`.
8. Débora C. Corrêa and Francisco Ap. Rodrigues. A survey on symbolic data-based music genre classification. *Expert Systems with Applications*, 60:190–210, oct 2016.
9. Pedro J Ponce De León and José Manuel Iñesta Quereda. Statistical description models for melody analysis and characterization. *International Computer Music Conference, ICMC 2004*, 2004.
10. Wei Chai and Barry Vercoe. Folk music classification using hidden markov models. In *Proc. of the International Conference on Artificial Intelligence IC-AI01*, volume 6. sn, 2001.
11. L. Knopoff and W. Hutchinson. An index of melodic activity. *Interface*, 7(4):205–229, dec 1978.
12. Umut Simsekli. Automatic music genre classification using bass lines. In *2010 20th International Conference on Pattern Recognition*, pages 4137–4140. IEEE, 2010.
13. Karsten Verbeurgt, Michael Dinolfo, and Mikhail Fayer. Extracting patterns in music for composition via markov chains. In *Proceedings of international conference on industrial, engineering and other applications of applied intelligent systems*, pages 1123–1132. Springer, 2004.
14. Wang Z.Y. *Analysis of Guqin collection (In Chinese)*, chapter 4, page 143. Central Conservatory of Music Press, 2005.

End-to-end Classification of Ballroom Dancing Music Using Machine Learning

Noémie Voss[1] and Phong Nguyen[2*],

[1] Sevenoaks School, Sevenoaks, United Kingdom
[2] Tokyo Techies, Tokyo, Japan
voss02@sevenoaksschool.org
phong.nguyen@tokyotechies.com
* The two authors contributed equally to this work

Abstract: 'Ballroom dancing' is a term used to designate a type of partnered dancing enjoyed both socially and competitively around the world. There are 10 different types of competitive ballroom dancing, each performed to different styles of music. However, there are currently no algorithms to help differentiate and classify pieces of music into their distinct dance types. This makes it difficult for beginner and amateur ballroom dancers to distinguish pieces of music, and know which type of dance corresponds to the music they are listening to. We proposed using an end-to-end machine learning approach to help classify music into different types with efficient and high accuracy. We evaluated four machine learning models and found that a Deep Neural Network with three hidden layers is the model with highest accuracy of 83%. As a result, ballroom dancers will have an easier method of distinguishing between specific types of ballroom dancing music.

Keywords: ballroom dancing, classification, deep neural network, machine learning

1 Introduction

Ballroom dancing is a term used to designate a type of partnered dancing enjoyed both socially and competitively in dance festivals such as the Blackpool Dance Festival [1]. These dances are split into distinct categories. 10 of these are categorised as competitive ballroom dancing; these include Chacha, Foxtrot, Jive, Paso Doble, Quickstep, Rumba, Samba, Tango, Viennese Waltz and Waltz.

Each of these dances corresponds to a specific type of music. Despite their variations, it can be challenging for the human ear to distinguish their characteristics and classify pieces of music into a specific dance category.

However, this process of discrimination can be greatly facilitated using machine learning. A model can be created to aid professional and amateur dancers to easily match a piece of music to its corresponding dance type using knowledge about these unique characteristics. This can be of use when it is necessary to find a song for a specific event, or when curious about which type of dance to dance to a song.

Machine learning is a popular technology nowadays, utilised to help build applications which are able to address a variety of similar problems. Classification models can be trained using machine learning, and applied to new data. Using a data-driven approach, a machine can be made to learn the characteristics of various types of ballroom dance music, and hence be able to classify them. This way, we can avoid having to handcraft rule-based processing for each individual dataset and dance type; rather, the machine can learn to classify music accurately and efficiently.

In this paper, we propose a method to classify ballroom music using a machine learning approach and evaluate different models, including Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Random Forest, and Deep Neural Network (DNN). We wish to embed this method into a mobile application to help beginners learn ballroom dancing more easily. With a public dataset consisting of 3992 pieces of audio data, we found that the classification accuracy of the models we tested are 60%, 50%, 76% and 83% respectively.

2 Related Work

In this section, we briefly describe similar, already existent, work using machine learning in music classification.

In recent years, machine learning has been applied widely for classification tasks, especially in the signal processing field. The most popular task is to classify music into different categories, one that has been studied since the early days of the internet. Tzanetakis and Cook (2002) addressed this problem with supervised machine learning approaches such as Gaussian Mixture model and k-Nearest-Neighbour classifiers [6]. Three sets of features utilised in this task were timbral structure, rhythmic content and pitch content. Support Vector Machines (SVM) were introduced later by Scaringella and Zoia (2005) [7]. With the recent success of Deep Neural Networks, a number of studies applying these techniques to speech and other forms of audio data were undertaken, such as by Gemmeke et al. (2017) [8]. However, representing audio in the time domain for input to neural networks is not very straight-forward, because of the high sampling rate of audio signals. Therefore, feature engineering for audio data is still a more popular technique for audio signals. H. Bahuleyan (2018) has extracted both frequency features and time-domain features, and used Deep Neural Network (DNN) to classify music genre with high accuracy [9].

U. Marchand, G. Peeters (2016) [10] have proposed a method to classify music into different ballroom dancing types by representing music data based on the application of Scale Transform along the two dimensions of time and frequency. However, this technique does not use a machine learning approach, but rule-based classification.

3 Music and Sound Data

Sound data consists of a wave converted into an electrical signal. Firstly, the sound wave is translated into an analog signal. An analog signal is a continuous representation of a sound wave, and can consist of any values. After being converted

into an analog signal, the signal is then converted into a digital signal using an analog-digital converter (ADC), allowing the sound to be represented in a way that can be stored digitally. Digital audio consists of a continuous sequence of numerical samples. A digital signal is formed by using the analog signal to capture digital values which represent the amplitude of the signal.

When working with sound waves, it is important to know that *frequency* is equal to the *pitch of the sound*. A higher frequency results in a higher pitch. Frequency is measured in Hz (hertz) units. The human hearing range is from 20-20kHz. To capture all frequencies that humans can hear in an audio signal, ADCs sample recordings at a frequency that is approximately double the human hearing range, at a rate of 44,100Hz. [2]

4 Features Engineering for Sound Data

4.1 Data Collection and Preprocessing

To create and train a model to develop an accurate method of matching a particular piece of music to a specific type of dance, an initial requirement is the collection of data.

We used a public dataset called the Extended Ballroom dataset [11] containing ballroom dancing music excerpts extracted from the website www.ballroomdancers.com. The dataset consists of 4180 tracks, corresponding to 13 different dance types.

Table 1. Number of tracks in Extended Ballroom dataset

Ballroom Dance Category	Number of audio tracks
Chacha	455
Jive	350
Samba	468
Viennese Waltz	252
Waltz	529
Tango	464
Rumba	470
Quickstep	497
Foxtrot	507
Paso Doble	53
Salsa	47
Slow Waltz	65
West Coast Swing	23
Total	**4180**

We decided to omit Salsa, Slow Waltz and West Coast Swing, because they are not performed competitively. We also omitted Paso Doble, as the dataset of 53 tracks

was not sufficient to train the model, and would have caused an imbalance. The final dataset consisted of 3992 tracks.

In order to increase the number of music samples for the machine learning algorithms, the data was cut into smaller segments, called 'windows': 10 seconds per piece, with 5 seconds of overlap. This also allowed the size of each piece of music to be standardised, making it easier to extract statistical features.

4.2 Feature Extraction

A total of 37 features, listed below, were extracted from each track. The first feature was the BPM (beats per minute, or tempo), which indicates the speed of the music. Each dance type has a specific range of tempos at which they are danced.

Table 2. BPM of different categories of ballroom dancing [3]

Ballroom Dance Category	BPM
Chacha	120-128
Jive	168-184
Samba	96-104
Viennese Waltz	174-180
Waltz	84-90
Tango	120-140
Quickstep	200-208
Rumba	100-108
Foxtrot	112-120
Paso Doble	120-124

The second feature extracted was the time signature (count). The time signature of a piece of music is equivalent to the number of beats per measure, indicating the rhythm of the music. The main significance of this feature is to facilitate the differentiation between Waltz/Viennese Waltz, which has a time signature of 3/4, and the other categories of dance music, which have a time signature of either 2/4 or 4/4.

A third feature extracted from the pieces of music was the fingerprint record of the song. A fingerprint consists of a list of five values combined to form one hashtag. These values are the *most common frequencies* of the piece of music from within different frequency ranges. In a sound bite, frequency is the determinant of pitch. Therefore, the fingerprint of a song can be interpreted as a series of pitches, which appear most frequently within their frequency ranges, within the song. The five ranges were: 30-40Hz, 41-80Hz, 81-120Hz, 121-180Hz, and 181-300Hz. [4]

The final 34 features extracted further increased the ability of the final model's classification. Both short-term and mid-term features were extracted from each piece of music. Short term features can be described as features extracted from the piece of music after having split it into several short-term windows. Mid-term features can be described as certain statistics extracted from the short-term features, such as the mean and the standard deviation.

Table 3. Features of the PyAudioAnalysis program [5]

Feature ID	Feature Name	Feature Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).

4.3 Classification Models

Waltz and Viennese Waltz have a different time signature to the other types of ballroom music, and therefore our classification model only uses time signatures as the main criteria to classify Waltz and Viennese Waltz from the other categories of dance. A second, separate, model is used to classify to which of the other seven types the song belongs to: Tango, Samba, Chacha, Rumba, Quickstep, Foxtrot Blues, or Jive.

The overall process (illustrated in Figure 1) results in a program able to classify music into the nine categories of ballroom dancing music.

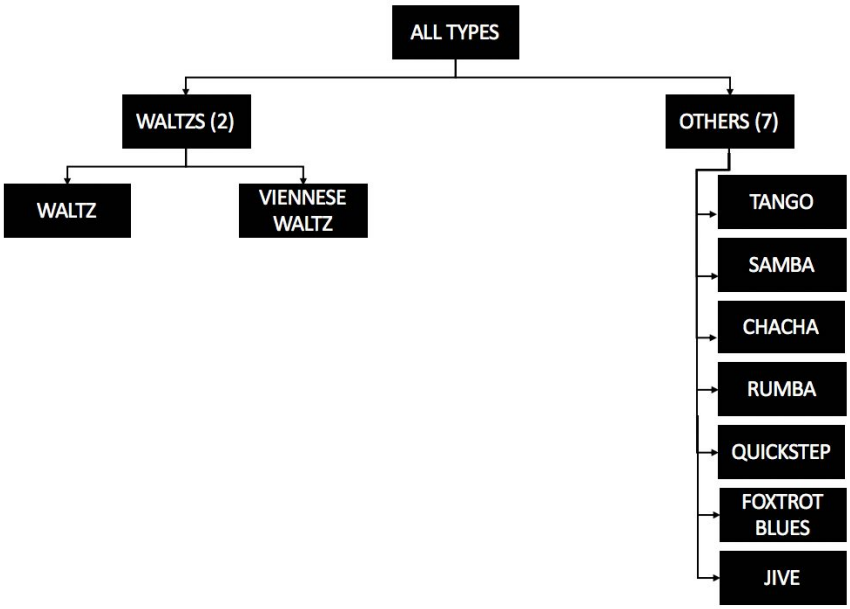


Fig. 1. Overview of the classification method.

To classify the other seven types of ballroom dance music and Waltz or Viennese Waltz, we compared seven types of classification models: Support Vector Machines (SVM), k-Nearest Neighbors, Random Forest and Deep Neural Network.

Support Vector Machines (SVM) is a machine learning algorithm used for classification. The SVM model looks at the training data and creates an optimal decision boundary, known as a 'hyperplane', between the extremes of the different classes of data. The hyperplane separates the data into the different categories. [12]

A k-Nearest Neighbors model classifies data based on a number (k) of its nearest neighboring data points, regardless of their label. The unidentified data point is categorised based on the label of the majority of its neighbors, i.e., it takes the same label as the label which appears the most. [13]

A Random Forest model uses decision tree algorithms to create several decision trees which can classify data into discrete categories. Each separate tree predicts a label for the unidentified data point. The final label is the label with the most predictions. [14]

A Deep Neural Network consists of machine learning algorithms similar to the human brain. These algorithms are able to recognise patterns based on the features of the data. Using hidden layers, the model is able to classify each of the data points. [15]

Figure 2 below describes how we processed data to train a machine learning model to classify different ballroom dance types, and evaluate the model.

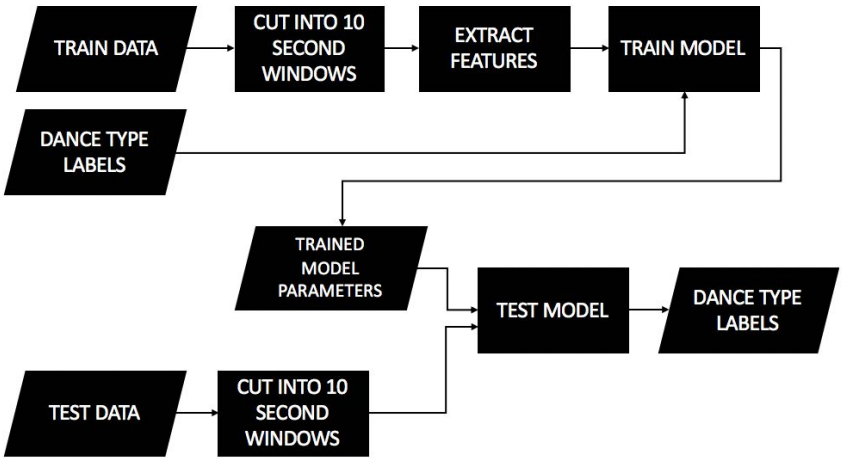


Fig. 2. Overview of the process.

4.4 Classification of a New Song

To use our method to classify a new song, we preprocessed the audio tracks by cutting them into 10-second windows with a five second overlap. For each standardized 10 second piece, after all features are extracted, we utilised time signature feature to group the pieces with a time signature of three as the first group (for Waltz and Viennese Waltz), and the second group of all other seven types of dance.

In the final output, each 10 second window piece belonging to a song had a classification label. We determined the ballroom dance label of the song based on the majority characterization.

5 Experimental Settings

The first step taken in creating the model was collecting the necessary data. We collected the data from the public Extended Ballroom dataset. As mentioned above, we removed the tracks for four types of dance: Paso Doble, Slow Waltz, Salsa and West Coast Swing. The created a final data set of 3992 tracks.

In order to evaluate our method, we split the data randomly with 10% of the data as the test data, and 90% of the data as the train data. In addition, we also tested the model on our own collected dataset with songs chosen by professional dancers. We picked 10 songs for each type of dance.

After the preprocessing, the total number of 10-second pieces was 45000 for the train data and 4100 for the test data. We then extracted the aforementioned 37 features from each piece and built four different classifiers: SVM, k-NN, Random Forest, and DNN. We evaluated each model in terms of accuracy and weighted F1-score. Hyperparameters of each model were tuned to their best performance by trial-error processes.

The models were not only evaluated on the test data, but also on a dataset of 30 professional ballroom dancing songs.

6 Results

Several different types of classification models were trained and evaluated in order to maximise the accuracy of the final model chosen. These models were: 'Support Vector Machine', 'K-Nearest-Neighbours' with neighbours=101, 'Random Forest' with 100 trees, and a 'Deep Neural Network' with three hidden layers.

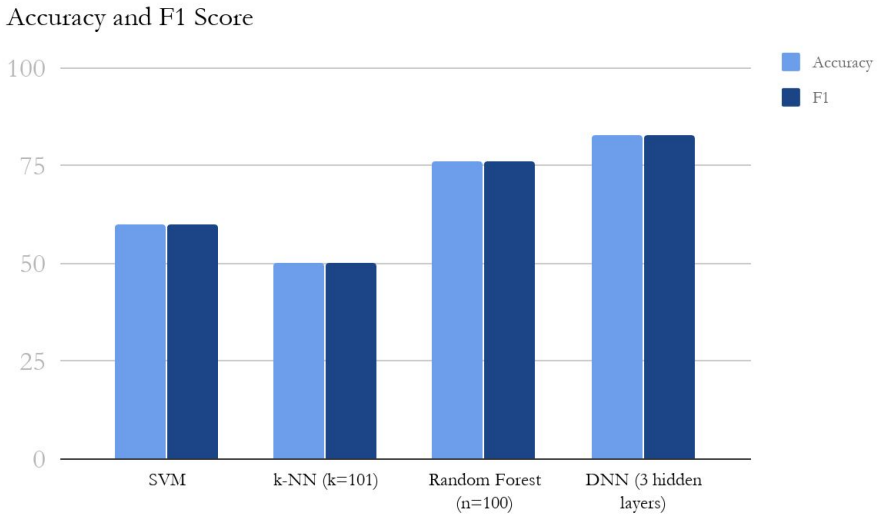


Fig. 3. A graph comparing the F1 scores and accuracies of each type of classification model.

The accuracy and F1 scores of the models are specified in Table 4.

Table 4. Accuracy and F1 scores of different classification models

Model	Accuracy	F1 Score
SVM	60%	60%
k-NN (n=101)	50%	50%
Random Forest (n=100)	76%	76%
DNN (3 hidden layers)	83%	83%

The results showed that the DNN model with 3 hidden layers has the highest accuracy. When we tested it on a dataset of 30 songs for ballroom dancing, it could recognize 27 songs accurately, which translated to 90% accuracy.

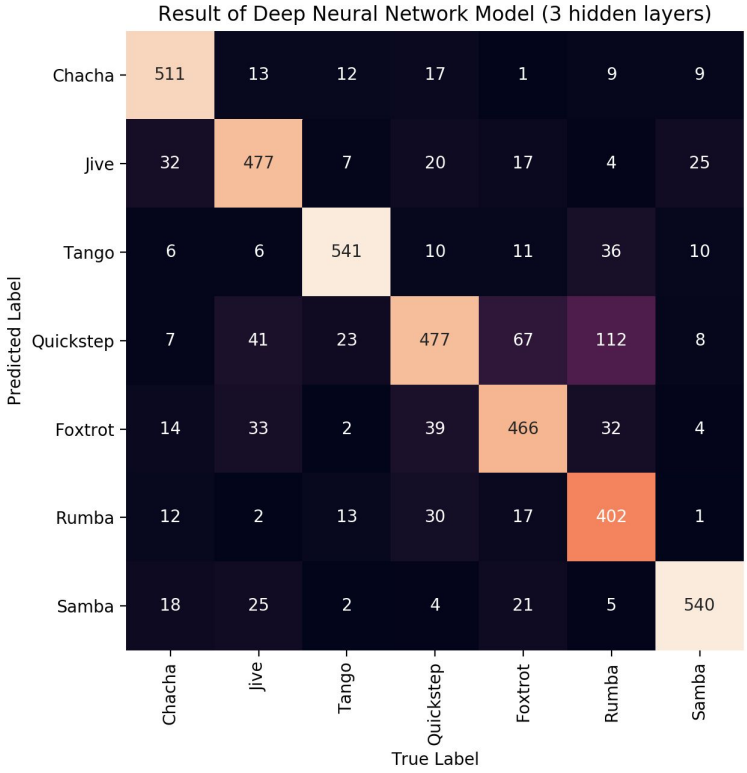


Fig. 4. Confusion matrix of Deep Neural Network model with three hidden layers for the 'Extended Ballroom' dataset.

Figure 4 above is a confusion matrix to show the results of the Deep Neural Network model with three hidden layers when tested on the Extended Ballroom dataset of 3992 tracks. The model is accurate at classifying both Tango and Chacha. However, the model is not able to classify Rumba and Quickstep as well as the other types of dance; 112 Rumba tracks were incorrectly labelled as Quickstep.

The overall accuracy of the model is 83%.

Once programmed into a mobile application, beginners in ballroom dancing will be able to use this application to improve their ability of distinguishing between various music, hence improving their confidence in ballroom dancing. This application can also be gamified to encourage beginners to guess the dance category before checking their answers with the model, providing a fun way for dancers to

improve their ability to distinguish between dance types. The actual mobile application has not been implemented yet, but could be feasible in the future.

7 Conclusion

We have evaluated an end-to-end method using machine learning to classify a piece of music belonging to a specific type of ballroom dancing. The Deep Neural Network with three hidden layers showed an accuracy of 83%. We open our source-code of the processing to train the model at:
https://bitbucket.org/xphongvn/ballroom_music_classification/

One way to improve this method will be finding more relevant sound features that can help classify ballroom dancing music with higher accuracy. An alternative way is to have more training data. We plan to publish a bigger dataset for ballroom dancing music in the future.

References

1. Blackpool Dance Festival, Website, Accessed in 2018 August, <http://http://www.blackpooldancefestival.com/>
2. Jovanovic, J, "How does Shazam work? Music Recognition Algorithms, Fingerprinting, and Processing", Website, Accessed in 2018 September, <http://https://www.toptal.com/algorithms/shazam-it-music-processing-fingerprinting-and-recognition>
3. Beats Per Minute Online, "Tempo Indications and Beats Per Minute (BPM) Reference for Social Dance Genres", Website, Accessed in 2018 June, <http://http://www.beatsperminuteonline.com/en/home/bpm-beats-per-minute-reference-for-dance-genres>
4. Jovanovic, J, "Music Recognition: Fingerprinting a Song", Website, Accessed in 2018 Sep, <http://http://https://www.toptal.com/algorithms/shazam-it-music-processing-fingerprinting-and-recognition>
5. Giannakopoulos, T, "3. Feature Extraction", Website, Accessed in 2018 August, <https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction>
6. George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing 10(5):293– 302.
7. Nicolas Scaringella and Giorgio Zoia. 2005. On the modeling of time information for automatic genre recognition systems in audio signals. In ISMIR. pages 666–671.
8. Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, pages 776–780.
9. Bahuleyan, H. (2018). Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149.
10. Marchand, U., & Peeters, G. (2016, September). Scale and shift invariant time/frequency representation using auditory statistics: Application to rhythm description. In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.

11. Marchand, U., & Peeters, G. (2016). The extended ballroom dataset.
12. Ray, Sunil, and Business Analytics. “Understanding Support Vector Machine Algorithm from Examples (along with Code).” Analytics Vidhya, 11 Mar. 2019. www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.
13. Srivastava, Tavish. “Introduction to KNN, K-Nearest Neighbors : Simplified.” Analytics Vidhya, 7 Mar. 2019, www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.
14. Koehrsen, Will. “Random Forest Simple Explanation.” Medium, Medium, 27 Dec. 2017, medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.
15. “A Beginner's Guide to Neural Networks and Deep Learning.” Skymind, skymind.ai/wiki/neural-network.

An evidence of the role of the cellists' postural movements in the score metric cohesion

Jocelyn Rozé[1], Mitsuko Aramaki[1], Richard Kronland-Martinet[1], and Sølvi Ystad[1]

Aix-Marseille Univ, CNRS, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France
roze@prism.cnrs.fr

Abstract. While playing expressively, cellists tend to produce postural movements, which seem to be part of their musical discourse. This article actually highlights the existence of a metric embodiment, i.e. a natural encoding of the score metric structure through specific periodicities of the musicians' postural movements. By the means of constrained postural conditions, we assess the degradation of the metric coupling between postural and bowing gestures among several cellists. Results reveal that bowing displacements should be in coherence with their postural behavior in order to ensure a correct encoding of the metric hierarchy.

Keywords: Cellist, ancillary/postural gestures, gesture-sound relationship, metric structure, musical expressivity

1 Introduction

Investigating musical gestures that are not directly related to the sound production is a research subject that has been given increased attention the last two decades. Often qualified of "accompanist" or *ancillary* to distinguish them from *instrumental* gestures directly responsible for the "effective" sound timbre [4, 1], these non-obvious performers' movements seem to play a role in the expressivity perceived by auditors, as well as the healthy nature of musician/instrument embodiment [8, 3]. According to the theory of embodied cognition [10], the whole musicians' body is involved during an artistic expression, and continuously shaped by three kinds of factors [23] : ergonomic (adaptation to the instrument), structural (adaptation to the elements written in the score) and interpretative (construction of a personal mental representation).

In this paper, we focus on the structural factor of embodiment, by assessing subtle mechanisms of synchronization between the cellists' ancillary gestures and the score metric structure. Such a *coarticulation* process [7] was demonstrated in previous studies through reproducible ancillary patterns localized at key points of the score for pianists [2, 20] or clarinetists [14, 6]. For bowed-string instruments, it was shown that coarticulation patterns also changed according to the type of bow stroke : short / *detached* or long / *legato* [24, 18, 22, 12]. Regarding the cellists in particular, we demonstrated in a previous work the alteration of short

rhythmic sections that turned out to be more salient in *detached* than in *legato* bowing mode [16].

Metric structure can be considered as a basis from which the musicians build their proper rhythmic perception. If this hierarchical organization isn't correctly integrated through suitable and distributed motor patterns, we may suppose that the cellists' musical sense of time collapses and leads to inexpressive performances [19]. As for the dancers actually, the intrinsic periodicities of cellists' motion patterns should present different levels of simultaneous synchronizations with the metric structure [21]. We can thus wonder if the slow movements of the cellists' chest or head would play a role in their natural expressiveness by intrinsically connecting to the musical time flow and the phrasing structure. Our study investigates the question by quantifying the phenomena of metric coupling between the ancillary and instrumental movements of cellists.

2 Protocol

2.1 Participants

The protocol described in this study is a subset of a larger experiment fully described in [16]. Seven cellists (4 males, 3 females) took part in this experiment. We chose musicians with high-level expertise, to ensure that any expressive degradations were due to postural constraints and not to technical weaknesses.

2.2 Apparatus

As we needed to analyze concurrently the motion and acoustic features produced by the cellists, an environment of two technological systems was set up :

- A eight-camera infrared motion-capture system (Vicon 8[1]). This system tracked the three-dimensional positions of 29 reflective markers positioned on the performers' body, and 9 additional markers placed on the cello and the bow at a frame rate of 125 Hz. The body markers were distributed according to the locations provided by the anatomical standard "Plug-in-Gait"[2], and corresponding to natural human joints or salient bone parts. A conversion to a simplified subset called *Dempster model* [5] was then carried out as an avatar of 20 markers corresponding to the 20 main body joints.
- A cello bridge pickup (Dpa 4096) connected to a MOTU interface (Ultralite MIC3) and configured at a frame rate of 44.1 KHz. The microphone location allowed to capture the acoustic features within the sound signal at source without being affected by potential reflections of the experimentation room. Audio and movement recordings were synchronized by means of a manual clap at the beginning of each recording.

[1] The Vicon 8 motion capture system used for the experiment was lent by the ISM (*Institut des Sciences du Mouvement)* of Marseille

[2] Vicon Motion Systems. Plug-in gait product guide. Oxford: Vicon Motion Systems, 2010, `http://www.irc-web.co.jp/vicon_web/news_bn/PIGManualver1.pdf`

2.3 Procedure

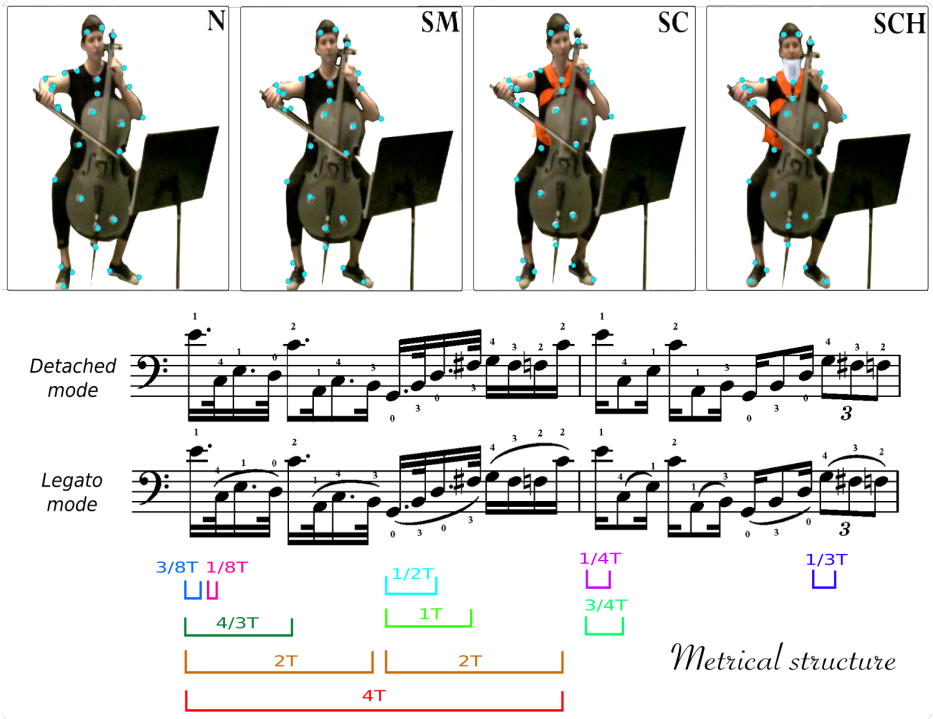


Fig. 1. The four postural conditions and the hierarchy of the score metric units

The cellists were asked to play a calibrated score at a slow tempo (♩= 45 bpm), while being subjected to four conditions of postural immobilization (Fig. 1).

- **N : *Normal condition***. Natural play as in a performance context.
- **SM : *Static Mental condition***. Stay as immobile as possible while playing.
- **SC : *Static Chest condition***. Physical semi-constrained situation, with the torso attached to the back of their chair by a 5-point safety race harness
- **SCH : *Static Chest and Head condition***. Fully constrained situation, with the torso attached as in the SC condition and a neck collar adjusted to limit head movements.

Whatever be the postural condition, they were informed to try playing in the most expressive way. The protocol also planned to assess the effect of the bowing mode (*detached/legato* for short/long bow strokes respectively), by proposing two versions of the same score with different bow stroke lengths. The cellists had

to achieve three repetitions of each combination of factors (postural condition and bowing mode) given in a random order for each participant.

The score was composed of six parts targeting specific technical cello difficulties. In this paper, we focus on a part with syncopated metric patterns that was assessed by the cellists as particularly difficult to perform in the constrained postural situations. The metric structure of this part, presented in Fig. 1, has been built as a grid of ten metric levels corresponding to relative time fractions of a quarter note beat (1T = 60/45 = 1.33s at ♩= 45 bpm). Our aim here was to establish relationships between this hierarchy of score metric units and the motion periodicities encoded at different levels of the cellists' body. To achieve that, we focused on the metric behavior of three marker trajectories located on the effective gesture side (mid-point of the bow) and on the accompanist gesture side (mid-points of the torso and of the head). We also simplified the analyses by only considering these marker displacements along the medio-lateral direction (cf Fig. 2), that was found as a prominent dimension of the whole cellists' movements [17].

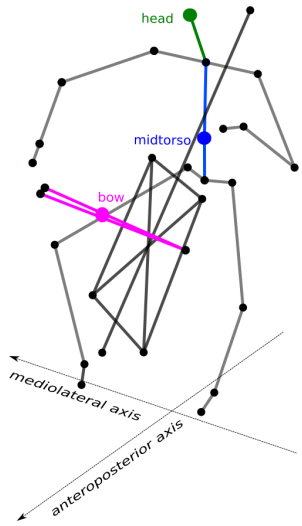


Fig. 2. The three markers considered for metric analyses

3 Analysis

3.1 Computation of a metric spectral centroid (MSC)

One way to quantify the overall metric embodied in a movement may consist in computing a centroid (or barycentre) of the metric units associated with

its periodicities. To do this, it is sufficient to extract a reduced version of the frequency spectrum of the movement, containing only the amplitude peaks that correspond to the metric units listed in the score (cf Fig. 1). From this reduced spectrum, we calculated the metric spectral centroid (MSC) by applying the same type of formula as the well-known spectral centroid descriptor[15, 9]. MSC descriptor thus stands for an average of the energies contained in the spectrum periodicities weighted by their metric rank:

$$MSC = \frac{\sum_{k=1}^{K} f_k A_k}{\sum_{k=1}^{K} A_k} \tag{1}$$

where k is the metric rank and K is the total number of metric units. For a given marker trajectory, the frequencies f_k and amplitudes A_k refer to its spectral peaks computed and associated as best as possible to a list of expected metric units.

In the context of our score part, we analyzed the metric of three marker motions (*bow*, *midtorso*, *head*) through a grid of $K = 10$ predefined metric units (cf Figs. 1 & 2). This process relied on four steps involving acoustic data (steps 1 & 2) and motion data (steps 3 & 4) :

1. Estimate the average duration of the quarter note beat (1T) in the sequence, based on the "Inter-Onset Intervals" (IOIs) of the notes that compose it. For each score bar, we computed a cumulated average of its note IOIs with respect to the four beats of the bar. This estimation of a bar quarter beat was then averaged among the four bars of the sequence to get a more accurate measure.
2. Build the metric hierarchy of the sequence by multiplying the estimated quarter note beat by the ten relative fractions of its metric grid. Next pass it to the inverse to get a metric grid of frequency values.
3. Conduct an FFT analysis (*Fast Fourier Transform*) to extract the most salient motion periodicities over the sequence duration. We chose to detect the 15 highest spectral peaks, as it empirically represented a good compromise between the total number of motion frequency peaks and the number of metric units to discover.
4. Select ten motion frequency peaks (among 15) likely to be the closest ones from the established metric hierarchy of ten frequency units. In this aim, we applied a two-way algorithm on a table associating the ten metric references (the keys) to the 15 identified spectral periodicities (the values). For each sequence of marker trajectories, the program solved an optimal distribution of ten periodicity values equal to a single motion frequency or zero in case of non-matching with the metric references.

Finally, the 10 key/value pairs of each distribution were used to get an estimation of motion metric centroid (cf Eq. 1). This MSC thus characterized a global motion synchronization on the most probable metric unit of the score part. The next two paragraphs illustrate this analysis process for sequences chosen among the cellists' repetitions in the bowing modes *detached* and *legato*. For

the sake of clarity, we only present them in the two opposite conditions (Normal and Fully constrained), that highlight the strongest metric differences (cf section 4).

3.2 Motion metrics for *detached* bowing mode

The Fig. 3 illustrates a representative tendency of motion metric analysis in *detached* bowing mode between the two opposite postural conditions (N and SCH) of a same cellist.

Regarding the motion trajectories of markers (Fig. 3.1), the periodicities of bow strokes seem to remain quite constant from normal to constrained conditions, whereas those of body ancillary parts (midtorso and head) switch from large patterns approximately matching the bar structure to faster oscillations synchronized with bow strokes. This suggests a natural synchronization of ancillary gestures on the slowest metrics of the score. The constraint would cause a breaking of this behavior as a stronger coupling with the instrumental gesture.

This tendency is confirmed by the motion spectral analyses (Fig. 3.2), with natural trunk periodicities (peaks at 0.18 Hz and 0.37 Hz) slowest than those of the bow (highest peak at 1.11 Hz). In the constrained situation, we actually notice a similar distribution of the bow metrics (highest peak at 1.17 Hz) in contrast with those of the midtorso and head that have increased a lot (highest peak at 1.76 Hz). Interestingly, the midtorso and head present a lot of common periodicities, but those of the head have globally higher amplitudes, suggesting a prominent role of the head in the handling of phrasing of wide musical patterns.

Such metric distributions even emerge more clearly when presented as reduced spectrums of 10 metric levels (Fig. 3.3) : From normal to constrained condition, the bow MSC remains localized on the metric level 3/4T, while the midtorso/head MSCs switch from low metric levels (2T and 4/3T) to high metric levels (3/4T and 1/2T). This sharp increase of three metrical units for both midtorso and head metric spectral centroids thus confirm a strongest coupling with the main bow metric level in situation of postural impairment.

3.3 Motion metrics for *legato* bowing mode

The Fig. 4 illustrates a representative tendency of motion metric analysis in *legato* bowing mode between the two opposite postural conditions (N and SCH) of the same cellist than previously.

Regarding the motion trajectories of markers (Fig. 4.1), the periodicities of body ancillary parts (midtorso and head) seem to match quite well those of the bow strokes in the normal condition, but desynchronize with much faster oscillations in the constrained condition. In contrast to the *detached* bowing mode, this suggests a uncoupling effect of the postural constraint, i.e. a natural accompanist and supportive role of the trunk movements for slow metrics of bow strokes, that became out of sync with the instrumental gesture in situation of postural impairment.

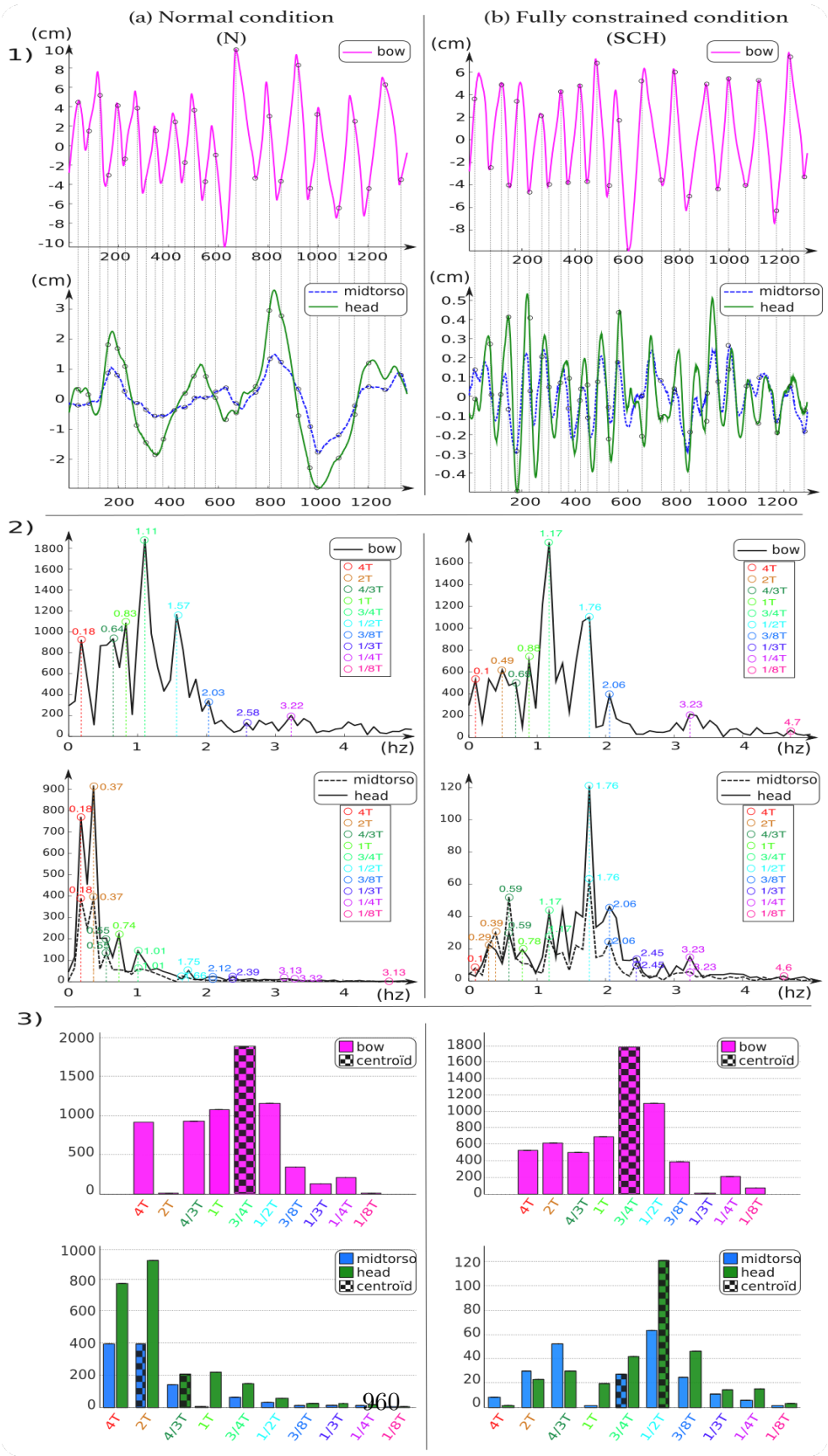


Fig. 3. Illustration of MSC computation for three markers of a cellist in ***detached*** bowing mode. **1)** Displacements along the medio-lateral direction (the vertical boundaries correspond to the IOIs of each note). **2)** Spectral analyses of these marker displacements (the 10 metric units of the score are matched to their detected periodicity peaks). **3)** Metric distribution and MSC localization of each marker between the two opposite postural conditions (N and SCH).

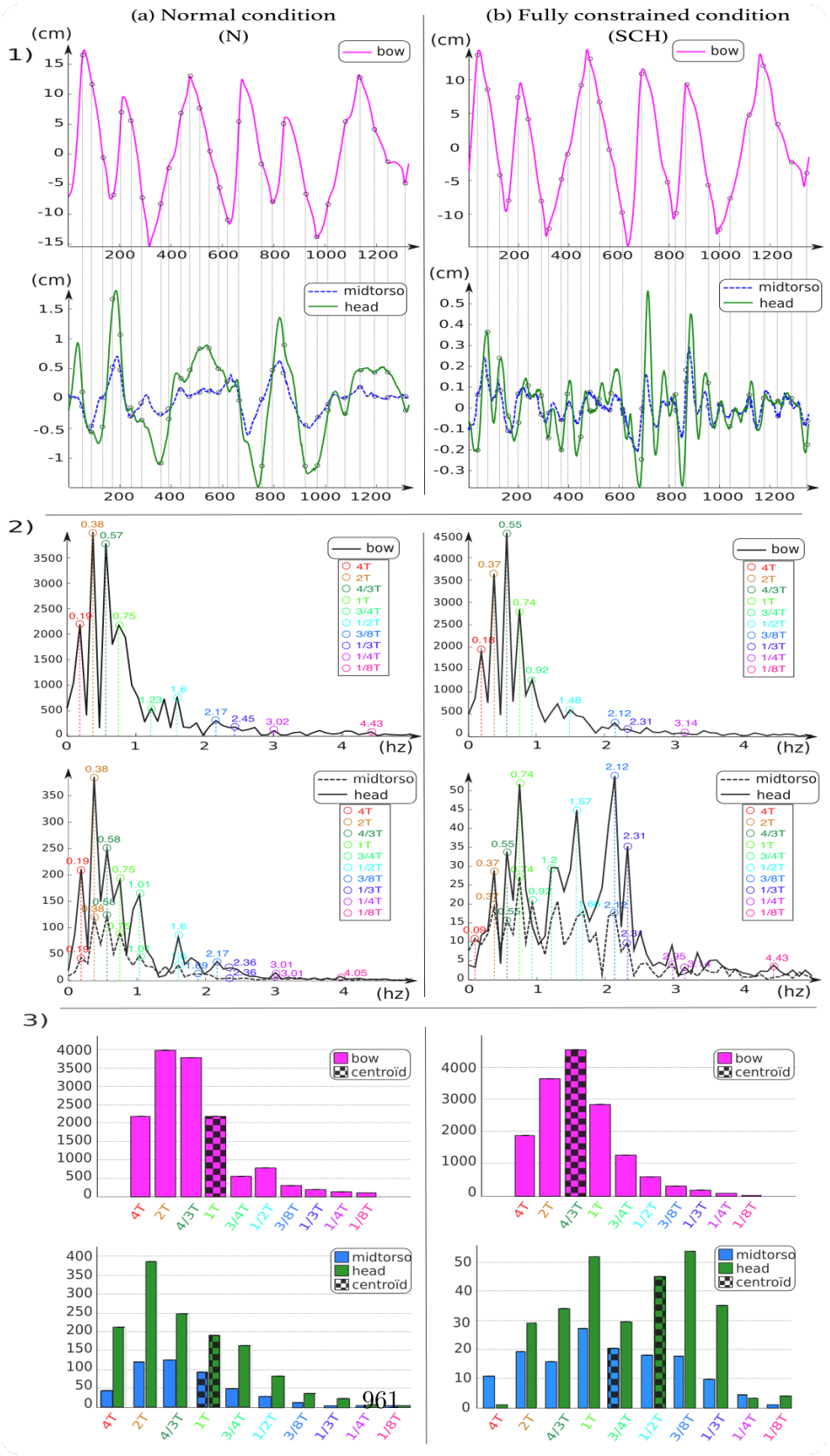


Fig. 4. Illustration of MSC computation for three markers of a cellist in ***legato*** bowing mode. **1)** Displacements along the medio-lateral direction (the vertical boundaries correspond to the IOIs of each note). **2)** Spectral analyses of these marker displacements (the 10 metric units of the score are matched to their detected periodicity peaks). **3)** Metric distribution and MSC localization of each marker between the two opposite postural conditions (N and SCH).

This tendency is confirmed by the motion spectral analyses (Fig. 4.2), with common periodicities between the trunk and the bow in normal condition : peaks at 4T (0.19 Hz), 2T (0.38 Hz) and 4/3T (0.5 Hz). In the constrained condition by contrast, the bow metric distribution seems to remain quite similar (highest peak at 0.5 Hz), whereas the postural encoding of the slowest score periodicities totally disappeared, increasing a lot towards smaller metric units : 1/2T (1.57 Hz) and 3/8T (2.12 Hz).

Such metric distributions even emerge more clearly when presented as reduced spectrums of 10 metric levels (Fig. 4.3) : Interestingly, they reveal that MSC localizations of the bow and trunk markers in the normal condition synchronized on the same metric level of a quarter note beat (1T). The constrained condition broke this natural metric coupling, since the bow MSC slightly decreased on the previous metric unit (4/3T), while the trunk MSCs increased on the next two units (3/4T and 1/2T for the chest and the head respectively).

4 Results

4.1 Generalization of the motion metric tendencies

To assess if the previous observations can be generalized to all the cellists, we conducted an analysis of variance (ANOVA) by repeated measurements over all their kinematic sequences. Such a technique is likely to quantify the inter-cellists stability of the metric centroid descriptor (MSC) of each marker through the two bowing modes and the four postural conditions. Simultaneously applied on the thee selected markers (*bow,midtorso,head*), this analysis becomes a multivariate anova (or MANOVA) of three dependent variables with two factors : [Postural condition] and [Bowing mode].

This MANOVA essentially reveals a significant effect of the postural condition factor ($F_{3,18} = 18.08; p < 0.05^{***}$) between the three MSC markers, regardless of the bowing mode. In Fig. 5 (*at left*) actually, we can first observe that for the natural condition, the MSCs of the three markers are metrically synchronized (between 0.9 and 1 Hz), in average on both bowing modes. Then, an interesting phenomenon occurs as the postural constraint increases: the MSC of the bow marker remains stable, while the MSCs of the trunk markers (chest and head) increase together, up to 1.4 Hz in the fully-constrained situation. Post-hoc LSD (Least-Square Differences) tests confirm significant MSC differences between the bow and the trunk markers for each of the three postural immobilization constraints. In the case of physical immobilization constraint by the chest alone, we also find a significant MSC difference between the torso and head markers.

The two-way repeated measures MANOVA didn't reveal significant MSC effects by bowing mode across the postural conditions. In Fig. 5 (*at right*) however, we can observe an interesting difference between the two bowing modes likely to confirm that the motion metric tendencies of section 3 can be generalized. Actually, the bow metric stability across the four postural conditions occurs around 1.2 Hz in *detached* bowing mode versus 0.75 Hz in *legato* bowing mode. As the

mean trunk metric increases in the same way whatever be the bowing mode, we globally get a situation of bow/trunk metric coupling (respectively uncoupling) in the *detached* (respectively *legato*) bowing mode as the postural constraints reinforce.

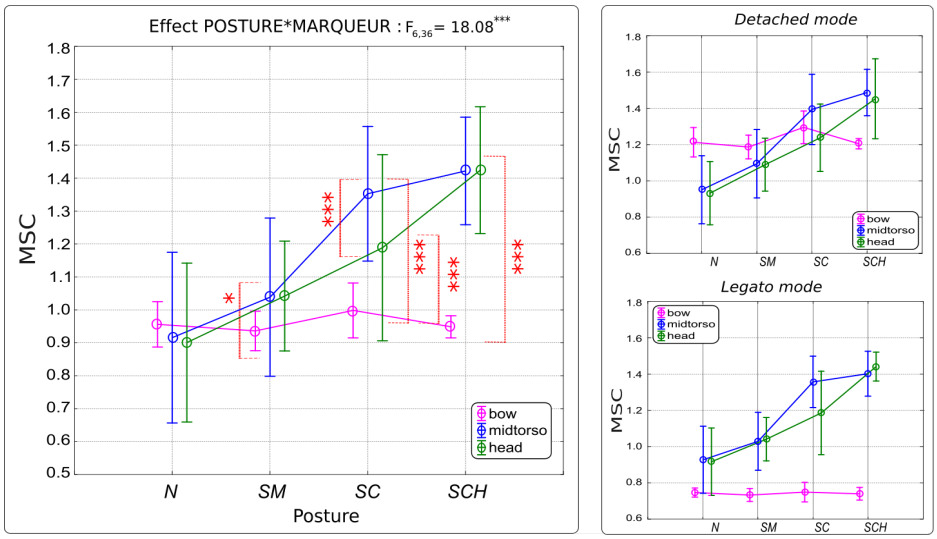


Fig. 5. MANOVA applied on the metric centroid descriptor (MSC) of three markers (*bow, midtorso, head*), according to the factors [Postural condition / Bowing mode] and performed by repeated measurements on the repetitions of the seven cellists

4.2 Discussion : the motion metrics cohesion

The results provided by MANOVA highlight a loss of the natural ***metric cohesion*** between the trunk and the bow motions as the postural constraints increase. Actually, the context of metric cohesion is reflected by the normal condition as a grouping of the metric mean values for the three markers (*bow, midtorso, head*). Then, a gradual desynchronization of the trunk metric with respect to the bow emerges from mental immobilization (SM), continues to expand by affecting more the chest than the head in partial immobilization (SC), until reaching a maximum in fully-constrained situation (SCH) for both chest and head. We here suggest that this phenomenon traduces a progressive disembodiment of the metric structure intrinsically encoded within the displacements of the cellists' segments : metric alterations would thus reflect different degrees of disorganization among the musicians' motor units.

Such results are consistent with the principles of embodied cognition, that predict different levels of metric encoding within the musicians' or dancers' body [21, 13, 11]. According to these principles, metric units of the musical structure

are encoded in body segments of proportional size : the trunk, that is a large segment, should encode the largest, i.e. the slowest, metric units. In contrast, the hand holding the bow is a small body extremity, that should encode smaller and faster metric units. Our experiment shows that the more the musicians' trunk is impaired in its displacements, the more it synchronizes with metric units too high (and too fast) for its natural behavior. Such an inconsistency of the cellists' motor units didn't prevent them from ensuring the bow metric required by the score, but the produced musical stream actually seemed more tense and mechanic, as played with a strict metronome pulse. This assumption is supported by the acoustic perception of the cellists themselves when interviewed about the loss of their natural sense of phrasing.

5 Conclusion

In this paper, we proposed a methodology to match the motion periodicities of the cellists' body segments with the metric units, pre-determined by the structure of the score. Postural constraints were used to assess the metric role of specific motor units such as the chest and the head, that indirectly contribute to the musicians' expressivity. Our analyses reveal that both of these ancillary body parts would play an important role to encode the slowest metric units of the score relating to musical phrasing. Impairing them actually resulted in losses of cohesion with the bow metric that reinforced with the postural constraint. The consistency and reproducibility of this phenomenon among the cellists allow to conclude that the musicians' postural flexibility is a key ingredient of their metric embodiment.

Acknowledgments. This work was partly supported by the French National Research Agency and is part of the "Sonimove" project (ANR-14-CE24-0018).

References

1. Cadoz, C., Wanderley, M.M.: Gesture-music : Trends in Gestural Control of Music, vol. 12. Ircam - Centre Pompidou (2000)
2. Davidson, J.W.: Qualitative insights into the use of expressive body movement in solo piano performance: a case study approach. Psychology of Music 35(3), 381–401 (2007)
3. De Alcantara, P.: Technique Alexander pour les musiciens. Editions AleXitère, Montauban, France (2000)
4. Delalande, F.: La gestique de gould: éléments pour une sémiologie du geste musical. Glenn Gould Pluriel pp. 85–111 (1988)
5. Dempster, W.T.: Space requirements of the seated operator: geometrical, kinematic, and mechanical aspects of the body, with special reference to the limbs. Tech. rep., University of Michigan (1955)
6. Desmet, F., Nijs, L., Demey, M., Lesaffre, M., Martens, J.P., Leman, M.: Assessing a clarinet player's performer gestures in relation to locally intended musical targets. Journal of New Music Research 41(1), 31–48 (2012)

7. Godøy, R.I.: Understanding coarticulation in musical experience. In: Proceedings of the International Symposium on Computer Music Modeling and Retrieval. pp. 535–547. Springer (2013)
8. Hoppenot, D.: Le violon intérieur. Van de Velde (1981)
9. Kim, H.G., Moreau, N., Sikora, T.: MPEG-7 audio and beyond: Audio content indexing and retrieval. John Wiley & Sons (2006)
10. Leman, M.: Embodied music cognition and mediation technology. Mit Press (2007)
11. Leman, M.: The role of embodiment in the perception of music. Empirical Musicology Review 9(3-4), 236–246 (2014)
12. Maes, P.J., Wanderley, M.M., Palmer, C.: The role of working memory in the temporal control of discrete and continuous movements. Experimental brain research 233(1), 263–273 (2015)
13. Palmer, C.: Music performance: Movement and coordination. The psychology of music p. 405 (2013)
14. Palmer, C., Koopmans, E., Carter, C., Loehr, J.D., Wanderley, M.M.: Synchronization of motion and timing in clarinet performance. In: Proceedings of the International symposium on performance science. pp. 1–6 (2009)
15. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the cuidado project. Tech. rep., IRCAM (2004)
16. Rozé, J., Aramaki, M., Kronland-Martinet, R., Voinier, T., Bourdin, C., Chadefaux, D., Ystad, S.: Assessing the influence of constraints on cellists' postural displacements and musical expressivity. In: Music, Mind and Embodiment - Postproceedings of CMMR, LNCS, vol. 9617, pp. 22–41. Springer (2016)
17. Rozé, J., Aramaki, M., Kronland-Martinet, R., Ystad, S.: Assessing the effects of a primary control impairment on the cellists' bowing gesture inducing harsh sounds. IEEE Access 6, 43683–43695 (2018)
18. Shan, G., Visentin, P.: A quantitative three-dimensional analysis of arm kinematics in violin performance. Medical problems of performing artists 18(1), 3–10 (2003)
19. Starker, J.: An Organized Method of String Playing. Bloomington, Indiana University Press (1975)
20. Thompson, M.R., Luck, G.: Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. Musicae Scientiae 16(1), 19–40 (2012)
21. Toiviainen, P., Luck, G., Thompson, M.R.: Embodied meter: hierarchical eigenmodes in music-induced movement. Music Perception: An Interdisciplinary Journal 28(1), 59–70 (2010)
22. Visi, F., Coorevits, E., Miranda, E., Leman, M.: Effects of different bow stroke styles on body movements of a viola player: an exploratory study. Ann Arbor, MI: Michigan Publishing, University of Michigan Library (2014)
23. Wanderley, M.M.: Quantitative analysis of non-obvious performer gestures. In: Gesture and sign language in human-computer interaction, pp. 241–253. Springer, Berlin, Heidelberg (2002)
24. Winold, H., Thelen, E.: Coordination and control in the bow arm movements of highly skilled cellists. Ecological Psychology 6(1), 1–31 (1994)

Zero-Emission Vehicles Sonification Strategy Based on Shepard-Risset Glissando

Sebastien DENJEAN[1,2], Richard KRONLAND-MARTINET[2], Vincent ROUSSARIE[1] and Sølvi YSTAD[2],

[1] Groupe PSA, Scientific Department and Disruptive Technologies, 78943 Vélizy-Villacoublay CEDEX, France
[2] Aix-Marseille Univ, CNRS, PRISM - UMR 7061, Marseille, France
sebastien.denjean@mpsa.com

Abstract. In this paper we present a sonification strategy developed for electric vehicles aiming to synthetize a new engine sound to enhance the driver's dynamic perception of his vehicle. We chose to mimic the internal combustion engine (ICE) noise by informing the driver through pitch variations. However, ICE noise pitch variations are correlated to the engine's rotations per minute (RPM) and its dynamics is covered within a limited vehicle speed range. In order to inform the driver with a significant pitch variation throughout the full vehicle speed range, we based our sonification strategy on the Shepard-Risset glissando. These illusory infinite ascending/descending sounds enable to represent accelerations with significant pitch variations for an unlimited range of speeds. In a way, we stay within the metaphor of ICE noise with unheard gearshifts. We tested this sonification strategy in a perceptual test in a driving simulator and showed that the mapping of this acoustical feedback affects the drivers' perception of vehicle dynamics.

Keywords: Sonification, automotive acoustics, multisensory perception

1 Introduction

With the increasingly tight limits of greenhouse gases emission imposed by governments, car manufacturers need to increase the share of electric motorizations in their series of vehicles. If the development of electric motorizations can help to reduce the in-use emissions of greenhouse gases, these quieter motorizations also lead to a radical change in the acoustic feedback that the driver perceives in the passenger compartment. This reduction of noise level in the passenger compartment has become a strong selling argument for such vehicles. However, although this reduction of noise is believed to be positive for the comfort of passenger, it can also become problematic in some cases. At low speeds for instance, when wind and rolling noises are very low, the loss of Internal Combustion Engine (ICE) feedback makes electric vehicles inaudible for pedestrians and can lead to dangerous situations [1-4]. Standards are now imposed to car manufacturers to equip their vehicles with Acoustic Vehicle Alert System (AVAS) to generate sounds that enable pedestrians to hear their vehicles.

In addition to the loss of informative noise outside of the vehicle, the loss of ICE noise is also an issue for the driver. ICE noise is indeed the most significant automotive noise source. On the one hand, it has a strong emotional power, being the main vector of evoked sportiness or powerfulness of the vehicle. Its loss can thus have a negative impact on driving pleasure and on brand identity. On the other hand, ICE noise is also the main acoustic information related to the vehicle dynamics that may influence driver's speed perception. Horswill and Plooy (2008) indeed showed that drivers underestimate their speed with a reduced acoustic feedback loudness. In a previous study, we also showed that acoustic feedback affects drivers' perception of acceleration (Denjean et al., 2013). The loss of ICE can thus affect the drivers' perception of both vehicle speed and acceleration.

With the development of electric motorizations, we loose a significant part of the dynamic information usually conveyed by ICE noise while conserving the disturbance due to wind and rolling noises. In a way, electric vehicles are not vehicles without noise but vehicles with only noise. We developed a sonification strategy to synthetize a new engine sound for electric vehicles to give back the information conveyed by engine noise in ICE cars.

2 Sonification Strategy

With the freedom of synthesized sounds in sound design, sonification of quiet(er) vehicles is a great opportunity to shape the soundscape of the vehicle. Sounds can give character to the vehicle, with a technological or even futuristic dimension. Sonification is a key for car manufacturers to work on its acoustic brand identity, and a tool to work on perceived quality of the cab acoustics. Sonification can also be used to inform the driver.

In this work, we focused on the dynamic feedback, setting esthetical design aside as a first step. We wanted this sound feedback to be able to inform the driver on the vehicle dynamics regardless of its design, and thus we worked on the invariants of sound dynamism. We chose to build our sonification strategy by mimicking ICE noise to achieve a natural and intuitive way of informing the drivers.

2.1 Objectives

We wanted our sonification strategy to convey information through the same levers as ICE noise, but with its own 'electric' personality. In a way, we wanted this sound to speak the same language as the ICE, but through a different voice.

Similarly, to ICE noise, we decided to inform the driver through pitch variations. However, pitch variations from ICE noise are correlated to the rotations per minute (RPM) of the engine and its dynamics is covered in a limited range of the vehicle speed. During acceleration, the pitch grows up to the high RPM of a given gear ratio before falling down at each gearshift. Thanks to this pitch decrease of each gear shift, the ICE pitch range is quite limited over the whole vehicle speed range. However, in electric vehicles, there are no gear shifts, and pitch range is therefore much bigger, since it is covered by only one gear. Even if we decided to simulate non-physical

acoustical gearshifts in our sonification strategy, we still would have to address the issue of the pitch dynamics in our sound. In ICE vehicle, the pitch of the engine noise grows around 2.5 octaves during one gearshift. It would not be possible to reproduce this pitch increase 5 times within the whole vehicle speed range. The sounds would simply be too high-pitched at high speeds. Even if we adapted the pitch range to fulfill the whole vehicle speed range, we would face too low pitch variations for low accelerations and provide information to the driver that would not be sufficiently precise to estimate his acceleration.

Sound synthesis was the solution to this issue, and an interesting answer to the problem was the Shepard-Risset glissando illusion with continually ascending or descending pitch. It could produce high pitch variations to precisely inform even for low vehicle accelerations while keeping a constrained pitch for the whole speed range of the vehicle. In a way, this illusion enabled us to conserve the metaphor of an ICE noise even with unheard gearshifts.

2.2 Shepard-Risset Glissando Based Sonification Strategy

Principle Shepard made tones that created the illusion of a musical scale that continuously ascends or descends in pitch [7]. These tones are composed of a superposition of sine waves separated by octaves with a loudness controlled by a raised cosine function. Low-pitched and high-pitched components are almost inaudible, while central components are clearly audible. These overlapping components are exactly one octave apart, and each scale fades in and out so that hearing the beginning or end of any given scale is impossible. When ascending this scale composed of twelve tones, we can hear the pitch growing at each tone. At the thirteenth tone, the first tone of the scale is heard again, but we still have the sensation of a growing pitch thanks to the specific construction of these tones. Risset used this illusion to create a glissando based on this scale, by simulating continuous transitions from one tone to the other.

We adapted this illusion to our sonification application. We chose to reduce the bandwidth of the raised cosine window to keep a limited range of octaves played at once. This modification allowed us to propose an additional control through the shifting of this window to change pitch through the spectral centroid of the generated sounds. To enhance the perceived illusion and open new design possibilities, we also chose to play chords instead of tones.

Synthesizer development We developed a synthesizer in Max/MSP allowing us to create these sounds and map our feedback to the vehicle speed and acceleration in real-time. Each parameter of the sine wave components can hereby be tuned and the raised cosine function can be used to define the loudness window.

Harmonic comb The user can choose the components of the chord selecting the tones on the musical scale. This leads to the creation of a harmonic comb composed by sine waves corresponding to these tones and duplicated at each octave. The frequencies of the components of this harmonic comb are controlled by a saw tooth function that browse the scale periodically. The amplitude of each component is defined by a loudness window.

Loudness window The user can define the bandwidth and the central frequency of the loudness window. This window shape responds to a raised cosine function and defines the amplitude of each harmonic component of the comb.

Mapping We mapped the pitch of the generated sounds to the vehicle speed. We chose a generic linear mapping between pitch and vehicle speed (S_v) based on the power law:

$$\text{Pitch} = kS_v + c \quad (1)$$

With the specific construction of our sound feedback, we have two independent ways to vary the pitch of the generated sounds: with the center frequency of the loudness window (F_c) and the sweeping speed (S_s) of the harmonic comb (the frequency of the saw tooth function). We declined our mapping to these two parameters, linking the logarithm of the center frequency of the window to vehicle speed, and deriving this formula to map the harmonic comb swiping speed to vehicle speed (S_v) and acceleration (A_v).

$$\log(F_c) = k_1 S_v + c_1 \quad (2)$$

$$S_s = k_2 A_v \quad (3)$$

Both controls induce pitch variations of the generated sounds and informs the driver about the vehicle speed (with the spectral centroid of the loudness window) and the vehicle acceleration (through the swiping speed of the harmonic comb).

We tested this sonification strategy in a perceptual test in a driving simulator to evaluate the impact of this mapping on the drivers' perception of vehicle dynamics.

3 Impact of Shepard-Risset Based Sonification Strategy on Drivers' Perception of Vehicle Speed and Acceleration

In the previous chapter we developed the sonification strategy to provide dynamic information through acoustic feedback to the driver. We focused our strategy on acoustic feedback on speed and acceleration, which is the main information used while driving and usually brought by ICE noise in cars with combustion engines. To evaluate the impact of the proposed sonification strategy on drivers' perception, we ran a perceptual test in a driving simulator.

3.1 Experiment Objectives

In a previous study we showed that the sound feedback plays a role in multisensory speed perception among drivers (…). The main goal of this experiment was to evaluate the impact of the sonification strategy on the drivers' speed perception.

However, speed also strongly depends on the preceding acceleration ((Salvatore, 1967; Recarte and Nunes, 1996). We designed our sonification strategy relying on

this information. The proposed sound, based on the Shepard-Risset glissando, is well suited to give strong acceleration information with the variation of pitch, but can give fuzzy information about absolute speed with its paradoxical pitch, even if we correlate the spectral centroid to vehicle speed.

We assumed that the sonification mapping would affect the drivers' perceived acceleration, and thus their speed production. The quicker the pitch variation the stronger the drivers perceived acceleration, inciting them to produce slower speeds.

To validate this assumption, we focused on the mapping between the swiping speed of the harmonic comb and the vehicle acceleration. We fixed the mapping between the filtering window displacement and the speed, and tested different mappings of this harmonic comb swiping speed according to acceleration.

We asked participants to accelerate or decelerate without the speedometer information to a provide a target speed in different sound conditions.

3.2 Method

Participants

29 volunteers (3 women and 26 men) employees of the PSA group, participated in this study. They all held a valid driver license and declared to have normal or corrected to normal vision and normal audition.

Experimental Device

We chose to run this perceptual experiment on a driving simulator to control the stimuli presented to the participants and ensure its repeatability. For this experiment we used the fixed-based driving simulator of the PSA group. This driving simulator is composed of the front half of a car disposed in front of a hemi-cylindrical screen. The scene used for the simulation represented a straight two-lane urban road. A picture of the driving simulator and an extraction of the scene are presented Figure 1.

Fig. 1. Picture of the driving simulator and the driving scene used for the perceptual study

The vehicle noise presented to the participants in the driving simulator was based on recordings in an actual car with a binaural dummy head. They were replayed in granular synthesis and our sonification proposals were mixed with this noise. This global mix was played back in the simulator thanks to a five loudspeaker sound system.

Experimental Variables

Vehicle speed and acceleration Participants had to reach 2 target speeds, 50 and 70 km/h, either accelerating from 30 km/h below the target (acceleration condition) or decelerating from 30 km/h above the target (deceleration condition). The acceleration was controlled and fixed to 2 values in the acceleration and deceleration conditions: +/- 0.75 m/s² or +/- 1.5 m/s². We thus tested 4 acceleration and 4 deceleration conditions.

Sonification parameters To focus on the impact of the harmonic comb speed variation on the perceived dynamics, all other sonification parameters were fixed. We used a major triad (tonal, major third and fifth) duplicated to other octaves as a basis for the sound feedback. This harmonic comb had been filtered with a 4 octave length window with a central frequency varying exponentially with vehicle speed from 120 Hz at idle to 660 Hz at 130 km/h (approximately 1 octave every 55 km/h).

For the mapping between swiping speed of the harmonic comb and the vehicle acceleration, we tested 3 linear mappings with constant k_2 = 0.04, 0.08 and 0.16. For the rest of the article, we will use the mapping with a constant k_2 =0.08 as reference and refer to the mappings as mapping 0.5 (k_2 = 0.04), mapping 1 (k_2 = 0.08) and mapping 2 (k_2 = 0.16). It is interesting to notice that we will encounter situations with the same sound feedback at different visual speeds, with mapping 1 at 0.75 m/s² and mapping 0.5 at 1.5 m/s², or mapping 1 at 1.5 m/s² and mapping 2 at 0.75 m/s².

We added a control condition without sonification to the 3 previous conditions.

Experimental variables summary These experimental variables that are summed up in the table have been crossed into a full factorial design. Each stimulus has been repeated 3 times, leading to an experimental design of 96 stimuli.

Table 1. Experimental variable summary. All variables were crossed in a complete experimental design and each stimulus was repeated 3 times leading to 96 stimuli by participant

Use case	Target speed	Acceleration level	Sound feedback
Acceleration	50 km/h	+/- 0.75 m/s²	Mapping 0.5 (k_2 = 0.04) Mapping 1 (k_2 = 0.08)
Deceleration	70 km/h	+/- 1.5 m/s²	Mapping 2 (k_2 = 0.16) Electric vehicle (no sonification)

Procedure

Participant tasks Participants had to perform a symmetric tasks during the acceleration and deceleration conditions (Fig. 2).

During acceleration, subjects were launched at a constant speed (20 or 40 km/h depending on the target speed 50 or 70 km/h). The driving simulation regulated the vehicle speed and the driver did not press the accelerator pedal, as if the subject drove with a speed regulator. When entering the acceleration zone, the subject pressed the accelerator pedal to accelerate. The vehicle acceleration was regulated by the simulation (at 0.75 or 1.5 m/s²) and the driver had to stop the acceleration by releasing the pressure on the accelerator when he thought he had reached the target speed. After the acceleration, the vehicle speed was kept constant by the driving simulation.

In the deceleration condition, the task was symmetrical, and the participants were asked to press the accelerator pedal to keep speed constant (at 80 or 100 km/h depending on the target speed). When they were ready to start, they reduced the pressure on the accelerator to begin the deceleration (at 0.75 or 1.5 m/s² depending on test condition) and had to press the accelerator again when they thought they had reached the target speed.

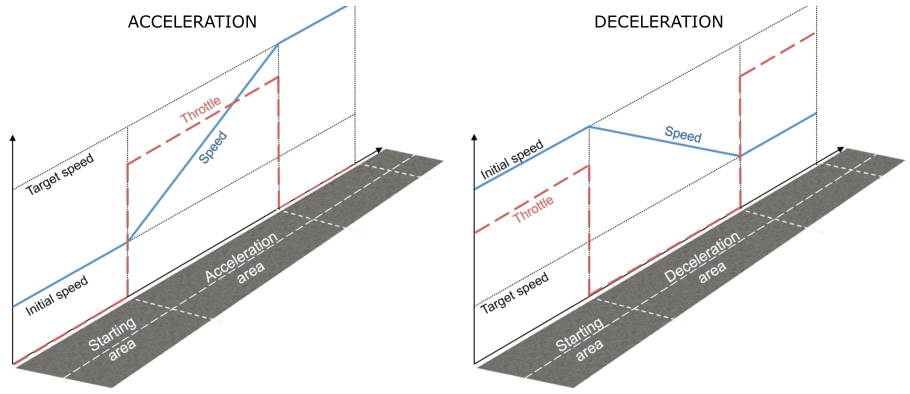


Fig. 2. Test procedure for acceleration and deceleration conditions

Test sessions Stimuli were divided in two blocs by the acceleration/deceleration condition. To avoid learning effects, the order of these sessions were pseudo-randomly launched between subjects. The whole test lasted approximately 1h45.

At the beginning of the test session, the experimenter presented the experiment to the participant. The experimenter explained that the participant would have to evaluate the vehicle speed without a speedometer feedback at different speeds. He also explained that different acoustic feedbacks would be provided, but without mentioning the nature or the type of control of these sounds.

Before the beginning of the test, the participants underwent a training phase. The first training phase aimed to familiarize the participant with the driving simulator: the participant drove freely in the map used for the test. During this phase, the speedometer was visible to the participant to calibrate his speed perception and no sonification was added. Participants were driving in the sound control condition (electric vehicle). When the participant felt comfortable with the driving simulator, the test began.

Each test bloc (acceleration and deceleration) began with a task-learning phase. The experimenter explained the subject's task and a couple of additional tasks were tested to make sure the participant fully understood the task. Stimuli corresponding to the different acceleration intensities, target speeds and sounds were presented with a pseudo-randomized experimental design based on a latin square to reduce learning effects.

3.3 Results

We recorded the speed reached by the participants after the acceleration/deceleration phase as their perceived target speed. We calculated the error between the speed reached by the driver and the target speed to compare the results between the two target speeds tested.

We ran an Analysis of Variance (ANOVA) on the whole set of data to determine the experiment parameters that had a significant impact on the speed reached by the drivers.

Acceleration condition In the case of acceleration, the ANOVA shows a significant effect of the sound condition ($p < 0.001$). A Scheffé post-hoc test separated these four condition in three groups: mapping 2 (mean error -1.6 km/h), mapping 1 (mean error -0.1 km/h) and electric condition (mean error 0.8 km/h), mapping 0.5 (mean error 1.1 km/h) and electric condition.

The intensity of the acceleration has also a significant impact on the speed reached by the drivers ($p < 0.001$), with a speed error of -2.8 km/h during acceleration and 3 km/h during deceleration.

We can also notice a significant effect of the target speed ($p < 0.001$) with a mean error of 0.9 km/h at 50 km/h and -0.7 km/h at 70 km/h.

Repetitions also showed significant effects ($p < 0.001$) with a mean speed error of -0.7 km/h for the first repetition, 0 km/h for the second and 1 km/h for the third repetition.

No interaction between these parameters reached significance level of 5 %.

Deceleration condition For the deceleration condition, the ANOVA shows the same significant effects. The sound condition has a significant effect on the speed reached by the drivers ($p < 0.001$) and Scheffé post-hoc test sorts the conditions in three groups: mapping 0.5 (mean error 8.5 km/h), mapping 1 (mean error 10.2 km/h) and electric condition (mean error 10.6 km/h), mapping 2 (mean error 12.4 km/h).

We can also notice a significant effect of the intensity of the acceleration ($p < 0.001$) with lower speeds reached at higher decelerations (mean error of 7.5 km/h at -1.5 m/s^2 and 13.3 km/h at -0.75 m/s^2).

The target speed has also a significant influence on the speed reached by the participants ($p < 0.001$), with a mean error of 10.8 km/h at 50 km/h and 10 km/h at 70 km/h.

As for the acceleration condition, repetition has also a significant influence on mean speed errors of 11 km/h for first, 10.6 km/h for the second and 9.6 km/h for the third repetition.

No interaction between these parameters reached significance level of 5%.

3.4 Discussion

Speed perception and influence of acceleration In Figure 3 we plotted the mean speeds reached by the participants in the acceleration and deceleration conditions for target speeds of 50 km/h and both acceleration intensities 0.75 and 1.5 m/s^2. We can notice a global overproduction of speed by the participants, suggesting that they

compensated their underestimation of speed by reaching speeds higher than the target. [9]. This underestimation of speed is higher at lower speed, in consistency with the literature [9-12], with a mean speed reached of 56 km/h for a target speed of 50 km/h and 77 km/h for a target speed of 70 km/h.

The value of speed reached by the participants is also modulated with the intensity and sign of the acceleration.

Regarding the sign of the acceleration, we can see that drivers tend to reach higher speeds during deceleration than acceleration, suggesting that they tend to overestimate their acceleration. This overestimation of acceleration makes them think that they accelerated (resp. decelerated) more than they actually did and stopped the acceleration (resp. deceleration) before they should, inducing them to reach lower (resp. higher) speeds.

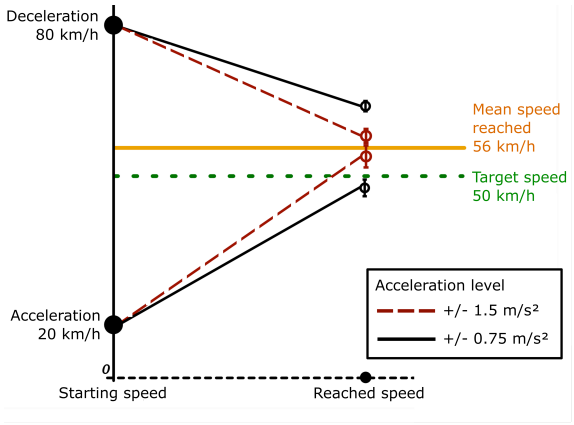


Fig. 3. Mean speeds reached by participants for the target speed of 50 km/h at acceleration and deceleration at both acceleration levels.

Regarding the intensity of the acceleration, we can notice that the situation is symmetrical. In the acceleration condition, the drivers tended to reach lower speeds at lower acceleration. Focusing on speed, we can infer that drivers underestimate their speed more after stronger accelerations, consistently with the study of Salvatore [8]. However, during deceleration, the opposite effect can be observed, with lower speeds reached with higher decelerations. This symmetrical result suggest that drivers tend to overestimate their acceleration and even more when its intensity its lower.

This result confirms the major role of acceleration in drivers' speed perception.

Influence of sound feedback Our sonification strategy based on the Shepard-Risset illusion is particularly well suited to inform on vehicle acceleration. More than the association of a pitch to a particular speed, this feedback enables to control the variation of the pitch according to vehicle acceleration.

We assumed that this feedback could modify the perceived acceleration among drivers and modify their speed production. The plot of speeds reached by participants in the different sound conditions confirms this hypothesis. Similarly, to the effect of the intensity of acceleration, we can notice that the effect of sound mapping in symmetrical during acceleration and deceleration. During acceleration, drivers tended

to reach lower speeds with a sound mapping that had a fast comb swiping. During deceleration, they reached higher speeds with fast comb swiping.

This result suggests that the sound information influenced the perceived acceleration among participants, with a perception of acceleration that increased with the comb swiping.

It is also interesting to observe the control condition without sonification which is included between the conditions with sonification. The comparison with the sonification conditions suggests than we can increase or decrease drivers' perceived acceleration according to the sound mapping.

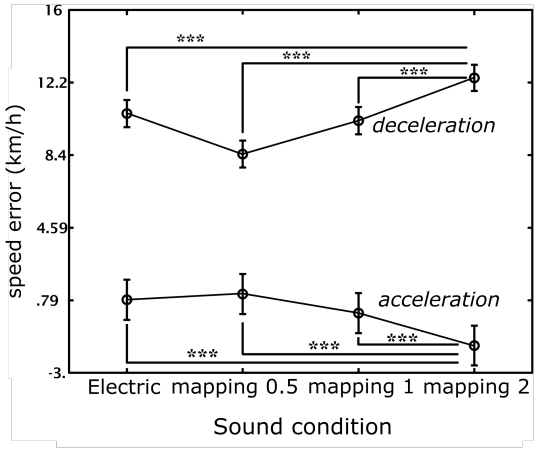


Fig. 4. Errors between speed reached by participants and target speed in acceleration and deceleration in the different sound conditions

Even if this effect is naturally lower than the effect of acceleration intensity, it can contribute to change drivers' acceleration perception. We saw that the drivers tend to overestimate their acceleration and mostly at low acceleration intensities. An exponential mapping can then be used to reduce this overestimation for low accelerations and adjust acceleration perception.

4. Conclusion

The development of quieter electric motorizations changes the soundscape perceived by drivers in their vehicles, and this "silence" is often praised as zero emission vehicles (ZEV). However, Internal Combustion Engine (ICE) noise is an important feedback for drivers. It gives precious information about vehicles' dynamics and has an important impact on the emotional level, since it contributes to the evocation of sportiness and fun when driving the vehicle. In this study we developed a sonification strategy to improve the pleasure of the driving experience in electric vehicles. In particular, we wanted the generated engine sound to provide information to the driver and to design the in-cab soundscape according to an emotional state. In this work we focused on the dynamic feedback of the vehicle. We decided to mainly focus on the

acceleration information, which is of major interest for driving control. The proposed strategy is based on the engine noise metaphor, with a pitch that increases with acceleration, to provide natural and intuitive feedback to the driver. We found that the Shepard-Risset glissando illusion was perfectly suited for this aim. It precisely provided information on the acceleration with fast pitch variations that could be maintained over the whole speed range of the vehicle. We tested different mappings of this sonification strategy with vehicle dynamics and showed that this had an impact on the drivers' acceleration perception and speed production. This sonification strategy can also be tuned to generate more technological sounds than ICE noise and is particularly well suited to answer the issue of interior sonification of zero emission vehicles (ZEV).

Acknowledgements.
This work was partly supported by the French National Research Agency (ANR-10-CORD-0003, MetaSon, "Métaphores sonores", https://metason.prism.cnrs.fr)

References

1. Garay-Vega, L., Pollard, J. K., Guthy, C., and Hastings, A.: Auditory detectability of hybrid electric vehicles by blind pedestrians. Transportation Research Record : Journal of the Transportation Research Board, 2248(1) :68–73 (2011)
2. Chamard, J.-C. and Roussarie, V.:Design of electric or hybrid vehicle alert sound system for pedestrian. Acoustics 2012 Nantes . (2012).
3. Ashmead, D. H., Grantham, D. W., Maloff, E. S., Hornsby, B., Nakamura, T., Davis, T. J., Pampel, F., and Rushing, E. G.: Auditory perception of motor vehicle travel paths. Human Factors (2012)
4. Altinsoy, E.: The detectability of conventional, hybrid and electric vehicle sounds by sighted, visually impaired and blind pedestrians. In Internoise 2013, Innsbruck (2013)
5. Horswill, M. S. and Plooy, A. M.: Auditory feedback influences perceived driving speeds. Perception, 37(7) :1037. (2008)
6. Denjean, S., Velay, J.-L., Kronland-Martinet, R., Roussarie, V., Sciabica, J.-F., and Ystad, S.:Are electric and hybrid vehicles too quiet for drivers ? In Internoise 2013, Innsbruck (2013)
7. Shepard, R. N.: Circularity in judgments of relative pitch. The Journal of the Acoustical Society of America, 36(12) :2346–2353 (1964)
8. Salvatore, S.: Estimation of vehicular velocity under time limitation and restricted conditions of observation. Technical report, Transportation Research Board (1967)
9. Recarte, M. A. and Nunes, L. M: Perception of speed in an automobile: Estimation and production. Journal of Experimental Psychology: Applied ; Journal of Experimental Psychology: Applied, 2(4) :291 (1996)
10. Evans, L.: Speed estimation from a moving automobile. Ergonomics, 13(2) :219–230 (1970)
11. Miloševic, S.: Perception of vehicle speed. Revija za psihologiju (1986)
12. Recarte, M. A., Conchillo, Á., and Nunes, L. M.: Traffic and transport psychocology, chapter Perception of speed and increments in cars, pages 73–84. Elsevier (2004)

"Tales From the Humanitat" A Multiplayer Online Co-Creation Environment

Geoffrey Edwards, Jocelyne Kiss and Juan Nino

[1] CIRRIS, Centre interdisciplinaire de recherche en réadaptation et intégration sociale
Laval University, QC, Canada
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
jocelyne.kiss@mus.ulaval.ca

Abstract. The fruit of more than three years of effort, "Tales From the Humanitat" is an online participatory environment developed to encourage various forms of co-creation in relation to a 19-book science fiction saga, written by Geoffrey Edwards and currently undergoing publication, and a range of production modalities. In particular, the environment was designed as a complement to a project aimed at developing a participatory opera based on one of the books of the saga. As a result, stations that provide interactive and participatory tools for music composition, dance choreography, vocal production and fan fiction elements are incorporated within the environment. The environment currently supports up to about 20 simultaneous participants, and allows for group co-creation initiatives to be organized. We will provide interested conference participants with access and coaching to use the environment.

Keywords: Cocreation, MMROPG, Virtual Participatory Opera.

1 Virtual Participatory Opera

The fruit of more than three years of effort, "Tales From the Humanitat" is an online participatory environment developed to encourage various forms of co-creation in relation to the ongoing development of the Ido Chronicles, a 19-book science fiction saga, written by Geoffrey Edwards currently undergoing publication and of which the first book is slated to be available next year (2), and other related modes of expression including a virtual participatory opera (3), a book of illustrations and exercises in fan fiction. Four forms of co-creation are supported - Musical composition, Dance choreography, Vocal production and sychronicity and Fan fiction - each at a distinct virtual station within the online environment. Furthermore, interaction is also orchestrated via a game scenario, encouraging players/participants to undertake activities at each of the four stations in order to gain access to a surprise oracular experience. The virtual environment presents the participant with aspects of the floating city of the Orr Enclave where the action of one of the books that makes up the Ido Chronicles occurs, along with the events taken up by the interactive opera based on the same volume.

We present in demo format this interactive multiplayer online environment. Interactive and participatory co-creation is seen as a way to engage contemporary audiences more fully in the production process and thereby ensure, in addition to greater engagement, better satisfaction on the part of audience members when engaging with artistic products. In the era of the mashup and user-generated content (4), not to mention social networks, new methods for engaging people are being sought (5).

Figure 1. The look of the online co-creation environment. Because the story takes place in an airborne city, structures are made out of fabric-like materials (that is, extremely light).

One way to achieve this is to solicit participation in the process that leads to the creation of new work. Not everyone wants to engage in this way, but there is a growing number of people who are interested. Furthermore, there is interest in diversifying the modalities of access to artistic performances and/or products. Our multiplayer online co-creation environment was designed in order to engage with potential audiences in this way, both in relation to live artistic performances and to support entirely online engagements.

The environment was designed in particular to complement our work on the development of participatory opera. Opera was chosen because it is one of the more complete forms of artistic performance (Novitz, 2001), incorporating elements from many different performance modalities - dance, choral music, solo lyrical performance, theatre and text, stage and costume design, and, potentially, also video and the graphical arts. Developing ways to engage audiences across the gamut of these diverse forms is a real challenge (6).

Rather than offering a completely open creative environment, for which supportive technologies and methods are still sorely lacking, our efforts were targeted to support our existing approach to participatory opera. Hence, for example, rather than offering users to develop and record new movements that could be incorporated into arbitrary dance sequences, we used the dance choreography we developed for our existing participatory opera, but provided users with the means to delay the dance sequence for different avatars. In this way, users may "play" with variations of the dance sequence we already digitized in order to generate new group dance arrangements. In subsequent steps we propose to offer a wider range of pre-digitized dance sequences that could be assigned to virtual dancers to explore group performances, as well as to potentially modify the dance movements themselves via interactive tools.

Figure 2. The four co-creation stations. (a) The Dance Monument ; (b) The Nemo Concourse; (c) The Jonah Agora; and (d) The Spiner Portal (fan fictions)

2 Public As Music Composer

For musical composition, we record the user's efforts to sing, and correct the pitch so that the sounds are "in tune", even if the original voice was not. These effects are presented to the user within the logic of the larger story line of the "Ido Chronicles". Hence the dance sequences are assigned to phantom presences called "phramae", while the sung fragments are associated with plants called "nemos". The nemos are generated by singing, and they grow faster in response to more intense efforts at vocal

production. In this way we provide real time visual feedback to the user concerning their use of voice. Once the melodic line is completed, the nemo replays the segment as long as an avatar remains in the vicinity of the plant. Since different players may each create their own nemos, there are opportunities to combined melodic lines across different participants or players. These are the same interaction modalities supported by the opera, and hence the online co-creation environment complements directly the opera performance.

Finally, and perhaps most interesting for musicians, are our efforts to support and encourage vocal production by participants. For this we draw on another creature within the science fiction universe that frames our online environment, the "jonahs". Jonahs are descendants of sperm whales that have been adapted to serve as interstellar transports. They communicate with humans by a kind of singing (as do contemporary whales). Our virtual environment records the breathing pattern of the user (via the microphone) and uses this to modulate the whale song produced by the jonahs. When the user's breathing pattern is synchronized with that of another user's (or the default pattern from a pre-recorded user), then the melody played is harmonious. When the two patterns are not synchronized, then the music produced is dissonant. Our idea is to enable connections between users that enhance their social interaction using a online form of non-verbal communication. This process allows non-musicians to take part of a duo.

3 CoCreation and Social Interaction

In the participatory opera we developed, we also used the breathing patterns of participants, measured via a wearable belt, to generate an atmosphere that supports flotation for creatures equipped with wings. This affected the unfolding story, supporting or opposing the main character's development, and hence allowed audience members to directly influence how the opera played out.

The fan-fiction component is, to date, the least developed component of the environment. We provide a kind of "portal" environment whereby participants may upload arbitrary documents to the site along with a "cover image" which provides clues as to the content of the document. These may then be accessed by other users.

The full online environment also offers a transportation system for moving between stations, again offered within the framework of the broader saga, although it is also possible to "walk" the avatars - a map-based navigation system is provided. There are also chat modalities provided, so that players may communicate with each other.

In order to engage with online users who might not be interested in the complementary opera project, but rather simply wish to explore the online co-creation environment, we created a framing "game". In each of the four stations, we hid a medallion

which the user must find and acquire. In addition to these four medallions, we hid a fifth medallion at another location. Once all five medallions are located, they can be activated at one of the transport stations, providing the user with access to a hidden oracle, itself based on texts within the 19-book saga. Furthermore, we created logon procedures and ensured that creations could be stored for later retrieval, as well as ensuring that the environment, along with its avatars, persists in time (and can be recreated if there is an interruption).

The environment, which was created within the Unity Game Engine, runs on a standalone server and currently supports up to about 20 simultaneous participants. It has been fully tested for roughly 40 active players, but performance drops sharply when the number of players exceeds 20. The environment is persistent, that is, it retains knowledge of which users are signed on, what avatars these control, as well as the melodies, dance arrangements, fan fiction documents uploaded, and jonah songs they have generated.

To access the environment fully, it is necessary to download an app which runs only in a Windows environment, and the logon procedure is somewhat messy. As a result, it has been hard to get buy in from potential users. We have therefore developed other access modalities that are lighter for the average user to implement (e.g. browser-based, and therefore not limited to any one operating system), hence providing co-creation experiences but not providing access to the full environment.

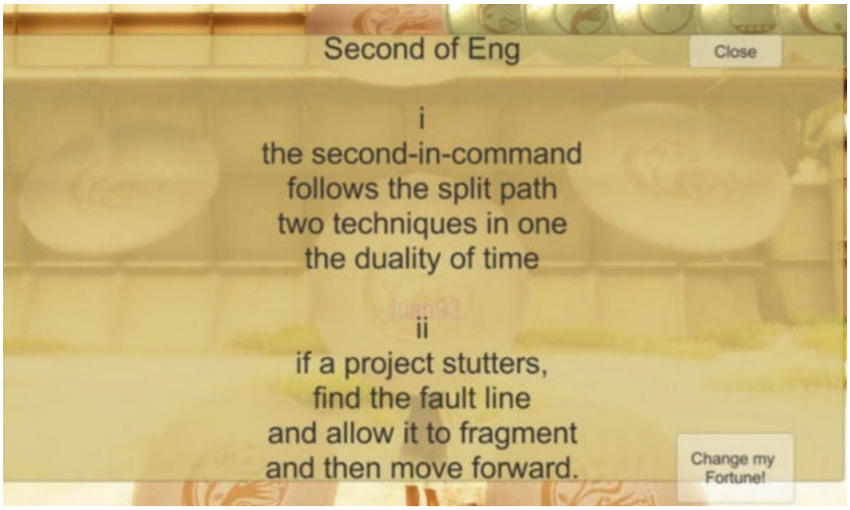


Figure 3. The Oracular System

The co-creation environment already created and operating is already being adapted and updated to serve other functions and projects that incorporate co-creation elements. For example, we are using the environment to help children with difficulties pronouncing certain sounds by giving them visual feedback. We also adapted the system to produce a 360 degree immersive experience for a collective interaction with the public. In this way we propose to engage the public in the creation of new drama by gathering their voices and their mutual energies.

The development of this online co-creation environment involved additional challenges compared to the development of the interactive opera itself (see Edwards et al., 2019, for more details concerning the latter effort). These involved designing and developing the 3D visuals, the avatars, and the navigational experience of online users moving through this complex environment (an airborne city of the far future), finding compelling interaction modalities consistent with the overarching storyline that frames this science fictional world, designing and testing the multiuser capability and interaction modes this creates, and testing the whole environment via a remote server. Although the opera project itself has been completed, the online environment continues to be used as a support for other projects.

References

1. Cheok, A.D., W. Weihua, X. Yang, S. Prince, F.S. Wan, M. Billinghurst et al. 2002. Interactive theatre experience in embodied+wearable mixed reality space. In : Proceedings of International Symposium on Mixed and Augmented Reality, ISMAR 2002, Darmstadt, Germany, 59-317.
2. Edwards, G. 2020. Plenum : The First Book of Deo. Boulder, Colorado, US : Untimely Books, in press.
3. Edwards, G., J. Kiss, E. Morales, C. McLaren, S. Lacasse, J.N. Falcon, J. Proulx Guimond and M.L. Bourbeau. 2019. Designing an Interactive and Participatory Opera, in Interactive Multimedia (ed. by D. Cvetkovic), Intech Publications.
4. Sonvilla-Weiss, S. 2010. Introduction : Mashups, remix practices and the recombination of existing digital content. In : Mashup Cultures, Heidelberg, Germany : Springer, 8-23.
5. Edwards, G., J. Kiss, E. Morales, C. McLaren, S. Lacasse, J.N. Falcon, J. Proulx Guimond and M.L. Bourbeau. 2019. Designing an Interactive and Participatory Opera, in Interactive Multimedia (ed. by D. Cvetkovic), Intech Publications.
6. Novitz, D. 2001. Participatory art and appreciative practice. The Journal of Aesthetics and Art Criticism. Vol. 59(2), 153-165.

Auditory Gestalt Formation for Exploring Dynamic Triggering Earthquakes

Masaki Matsubara[1], Yota Morimoto[2], and Takahiko Uchide

[1] University of Tsukuba
[2] mdoos
[3] National Institute of Advanced Industrial Science and Technology (AIST)
masaki@slis.tsukuba.ac.jp

Abstract. In seismology, seismologists usually visualize seismograms by plotting them on as a function of time. Proper visualizations help them determine the nature of earthquake source processes. However it is hard to determine the time span and frequency band to be focused on beforehand. In order to help investigating seismic wave propagation especially for discovering dynamic triggering earthquakes, we have been proposing interactive sonification system. The system leverages a capability of human auditory scene analysis to generate sound to be segregated into different objects. As a preliminary result, the seismic sonification for the 2011 Tohoku-oki earthquake successfully revealed a dynamic triggering event in the Hida area, Central Japan. The sonified sounds formed auditory gestalt and showed some characteristics and distributions such that seismologists can easily determine the time span and frequency band to be focused on. In this demo, we will provide the opportunity to listen sonified sounds and demonstrate the interactive sonification process.

Keywords: Interactive Sonification, Human Computation

1 Introduction

Seismologists usually visualize seismograms by plotting them on as a function of time. Sometimes they plot seismograms from multiple stations and filter them in order to emphasize the feature of the data. Proper visualizations help them determine the nature of earthquake source processes and/or the effects of underground structures through which the seismic wave propagates. However, it is hard to determine the time span and frequency band to be focused on only with visualization.

In order to help investigating seismic wave propagation especially for discovering dynamic triggering earthquakes, we have been proposing interactive sonification system [1–4]. The system leverages a capability of human auditory scene analysis [5], particularly *a law of common fate* in gestalt principle: i.e. Elements with the same moving direction are perceived as a single unit. Since seismic waves from neighbouring stations are correlated, the system sonifies sounds to be heard in the same stream. Thus when salient unexpected events occurs, the

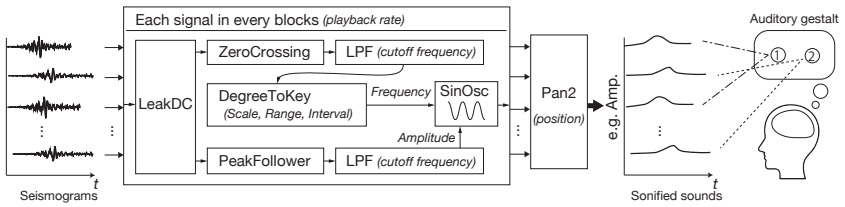


Fig. 1. System architecture. Each box represents common functions of SuperCollider, and italic words represent parameter variables.

sonified sounds can be easily segregated into different objects. The system also allows users to limit areas of stations for the sonification so that they can easily identify where a dynamic triggering event may or may not occur.

As a preliminary result, the seismic sonification for the 2011 Tohoku-oki earthquake successfully revealed a dynamic triggering event in the Hida area, Central Japan. The sonified sounds formed auditory gestalt and showed some characteristics and distributions such that seismologists can easily determine the time span and frequency band to be focused on.

For this demo, we have updated the system to the version 2.0. The contributions of the demo are as follows:

- Availability for the common data format among seismologists.
- Process improvement: Loudness correction, Speedup by parallel processing.
- GUI improvement: Temporal correction, Console and oscilloscope.
- New case study: Applied to seismograms recorded in Hokkaido.

In the demo, we explain the system architecture, provide the opportunity to listen sonified sounds, and demonstrate the interactive sonification process.

2 System Description

SoS (Sonification of Seismograms) is a sonification system of seismograms which started as a collaborative study among a data scientist, a seismologist and a composer [1]. We have built the sonification system using the SuperCollider sound synthesis programming environment as shown in Fig 1. As of version 2.0, the system is able to read the common data format among seismologists: SAC (Seismic Analysis Code) binary data which records seismograms in IEEE 32-bit floats. We have also made a number of improvements since the previous version which will be detailed in the following section. There are two main sonification methods which we have implemented: audification (direct playback of seismic data) and sonification using pitch and amplitude tracking. The practical use of the system is the following: we first read the SAC binary data, remove its DC offset (if present) and convert the data into audio files (using the implemented functions invoked from in the CUI commands); we then use the GUI to select a range of more than 500 stations to audify / sonify.

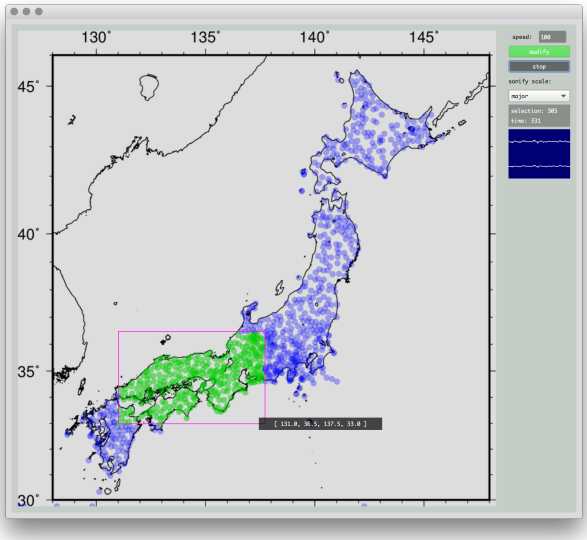


Fig. 2. The GUI Interface of SoS (Sonification of Seismograms) ver. 2.0

2.1 The GUI and improvements

As of version 2.0, we made many improvements on the GUI. The user can now select a range of stations using the mouse (Fig. 2) and the selected stations become highlighted (with a report of the number of selected stations). We have also added an oscilloscope which visualizes the waveform of the sonified sound. The user can specify the speed of audification / sonification using the number box. In case of audification the change in speed results in the transposition of the pitch itself due to the nature of the method. For the sonification method, however, the pitch structure is independent of the playback rate so that we can optimize the frequency range of sound taking into account the psychoacoustics of human audibility. The user can also select the pitch structure of sonification from a number of musical scales.

3 Application to Seismic Data

We applied SoS to seismograms of the 2011 Tohoku-oki, Japan earthquake (magnitude (M; Richter scale) 9.0) from the strong-motion seismic network, K-NET and KiK-net maintained by National Research Institute for Earth Science and Disaster Resilience (NIED) [6]. This earthquake is one of most significant earthquakes well recorded by modern seismic networks and brought substantial earthquake and tsunami disaster. The sonified sound of seismic data from 116 stations was bubble-like sounds as a result. The playback rate is 10. The sound gives an impression that the seismic waves were spreading out from the source.

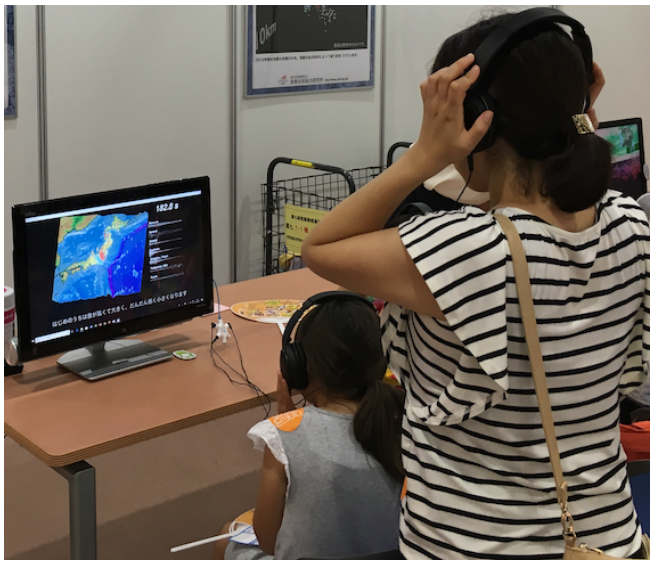


Fig. 3. Display of the sonified sound together with the visualization, in AIST Tsukuba Center Open House

An interesting sound was found around 230 s (in the original data; 23 s in the sonified sound) from the origin time. By sonifying the data from several areas using the SoS software, we identified that this sound was from Hida area, Gifu Prefecture, Central Japan. In seismological community, this was known as a dynamic triggering earthquake which was triggered by seismic waves from distant large earthquakes.

We have provided opportunities for people to listen to this sound in the AIST Tsukuba Center Open House every summer (Fig. 3). For most of them, it seems to succeed in feeling the seismic wave propagation and understanding the existence of the dynamic triggering event. The sonified sound gives a good chance to talk about those seismological phenomena.

From the seismological aspect, it is important to confirm if our method would be useful for detecting dynamic triggering earthquakes also for other major mainshocks. Before applying SoS to other large earthquakes, we sonified seismograms recorded in Hokkaido from individual small earthquakes that occurred also in Hokkaido: M 2.0, M 3.0, and M 4.0. This time we used seismograms recorded by short-period seismometers at Hi-net stations [7]. Note that, in the dynamic triggering cases, seismic waves from the mainshock and triggered events (usually small) are superimposed, therefore the single small earthquake case must be easier to hear. At this moment, it is difficult to find the M 2.0 earthquake in the sonified sound, while we can hear M 3.0 and M 4.0 earthquakes. This kind of experiments will be useful to identify the limitation in the exploration of earthquakes by seismic sonifications.

4 Conclusion & Future Work

In this demo, we explained SoS system ver. 2.0, which leverages a human capability of forming auditory gestalt and sonified sounds to be segregated for dynamic triggering earthquake discovering. As a preliminary result, the seismic sonification for the 2011 Tohoku-oki earthquake successfully revealed a dynamic triggering event in the Hida area, Central Japan. We applied further seismograms recorded in Hokkaido, and we could hear other individual M 3.0 and M 4.0 earthquakes, but not for M 2.0 one.

For future work, we will investigate semi-automatic parameter adjustment for efficient dynamic triggering earthquake capturing. We also plan to publish SoS system to seismologists community.

Acknowledgments. We used the Global CMT Catalog and seismic data from Hi-net of NIED, seismic data from JMA, Hokkaido University, Hirosaki University, Tohoku University, the University of Tokyo, Nagoya University, Kyoto University, Kochi University, Kyushu University, Kagoshima University, and AIST. This work was supported by JSPS KAKENHI 16H01744 and 17K14386.

References

1. Matsubara, M., Morimoto, Y., and Uchide, T. (2016). Collaborative study of interactive seismic array sonification for data exploration and public outreach activities. In Proceedings of ISon 2016, 5th Interactive Sonification Workshop (pp. 56-60).
2. Uchide, T., Morimoto, Y., and Matsubara, M. (2016). Seismic audification and sonification for data exploration, JpGU Annual Meeting, SSS28-P02.
3. Uchide, T., Morimoto, Y., and Matsubara, M. (2018). Sonification of seismograms for exploring dynamic triggering earthquakes, JpGU Annual Meeting, SSS10-P08.
4. Uchide, T, Morimoto, Y., and Matsubara, M. (2019). Dynamic triggering earthquakes exploration by sonification of seismograms, JpGU Annual Meeting, SSS11-P09.
5. Bregman, A. S. (1994). Auditory scene analysis: The perceptual organization of sound. MIT press.
6. National Research Institute for Earth Science and Disaster Resilience (2019b), NIED K-NET and KiK-net, National Research Institute for Earth Science and Disaster Resilience, doi:10.17598/NIED.0004.
7. National Research Institute for Earth Science and Disaster Resilience (2019a), NIED Hi-net, National Research Institute for Earth Science and Disaster Resilience, doi:10.17598/NIED.0003.

Minimal Reduction

Christopher Chraca

University of Victoria, Victoria BC V8P5C2, Canada
christopherchraca@gmail.com

Abstract. *Minimal Reduction* is an interactive sound installation where individuals will be put in the Engineer's Situation. Within this situation an individual will listen and manipulate the sound traits of recorded audio. This environment should lead an individual to enter the active listening axis which has reduced listening and heightened listening as the endpoints of the axis. This interactive sound installation is a demonstration of the Engineer's Situation and is meant to express the validity of the term. *Minimal Reduction* uses a Max/MSP patch to send audio out an Ambisonics surround sound system. The patch connects to an interactive device which has fourteen potentiometers, six knobs and eight faders, for individuals to control five sound traits of the recorded audio. The audio itself is an electroacoustic composition based on the musique concrète genre. Individuals who are on the active listening axis near reduced listening can understand the sounds or music by following the changes to the sound traits.

Keywords: Sound installation, Real time interaction, Schaefferian theory, Critical listening, Physical interface.

1 Introduction

Reduced listening is a concept for listening that stems from Pierre Schaeffer's final summary of listening intentions within the "Treatise on Musical Objects" [4]. The definition for reduced listening from the New Media Dictionary is "listening to sound for its own sake, in order to grasp its values and its character, without taking into account its source, what it reveals or its possible significance" [3]. Reduced listening marks the starting point of work and research that culminated in *Minimal Reduction*, but further exploration into post-Schaefferian critics as well as an audio engineering term, critical listening. Through this research came a new term the Engineer's Situation. The Engineer's Situation is how *Minimal Reduction* functions conceptually. The installation is an interactive device which is connected to a Max/MSP patch. This patch augments five distinct recorded sound objects that are then sent out an Ambisonics surround sound system. The outcome of this installation is to display the Engineer's Situation.

2 The Engineer's Situation

Reduced listening and critical listening require listeners to analyze the *sound traits*. These sound traits of recorded audio are frequency, amplitude, spectrum, spatial, and duration and time. Reduced listening involves passive listening that focuses on eliminating causal and semantic meaning while listening to recorded audio, while critical listening serves to augment the sound traits. Furthermore, critical listening does not distance itself from the causal and semantic meaning behind the sound source, but the engineer uses this information to help inform the augmentations to the sound traits.

Within the discourse of post-Schaefferian critique, Leigh Landy introduces the notion of an active listening axis that has reduced listening and heightened listening as its endpoints. Heightened listening [2] is the opposite of reduced listening, where listeners focus on the semantic and causal meaning of the recorded sound. Many post-Schaefferian critics have shown that it is almost impossible to achieve a pure form of reduced listening, so this axis may be used to deviate `from reduced listening towards heightened listening. Akin to how an engineer may use the semantic and causal meaning of a sound to inform their decisions, this axis allows individuals to indulge in semantic and causal listening.

Michel Chion stated, the instrumentalist who continuously listens and refines the sounds they create will enter a rudimentary reduced listening mode to focus on the traits of the instrument's sound [1, 4]. This rudimentary mode may be a far cry from pure reduced listening, but instead it should be part of the active listening axis. This same concept can be applied to recording engineers and their work, entering the active listening axis.

2.1 A New Term

The Engineer's Situation combines critical listening with the critiques of reduced listening by post-Schaefferian critics. Within this situation, an individual must interact and augment the sound traits of the recorded audio as they listen. The engineer mixing in a recording studio is a prime example of this situation in action. Fig. 2 shows a flowchart of the Engineer's Situation.

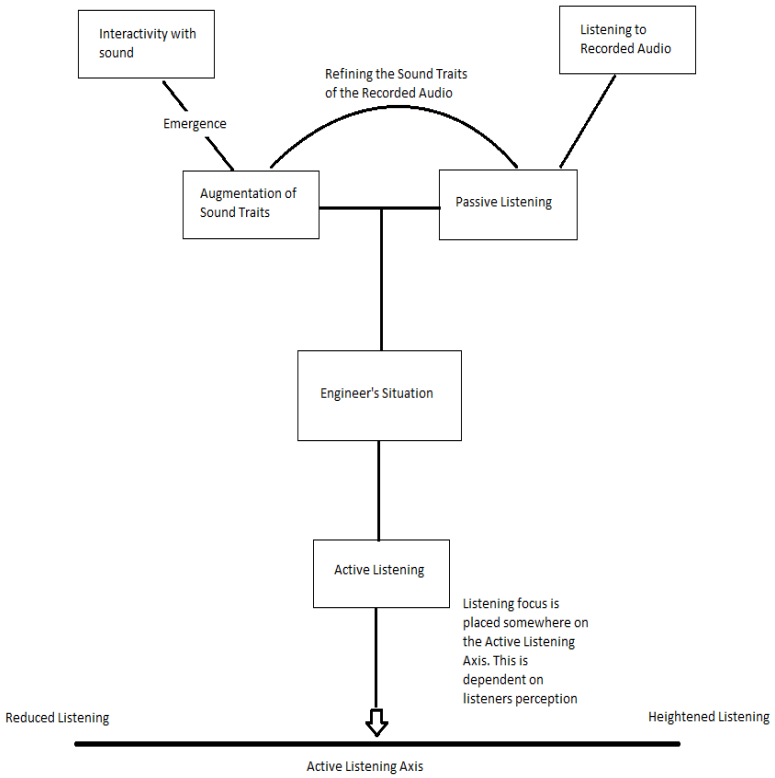


Fig. 1. Flowchart of the Engineer's Situation

3 Interactive Sound Installation Design

Minimal Reduction consists of a multi-speaker surround sound system around a table with the interactive device on top of it. The number of speakers can vary, but the set-up requires at least four. The visual design of this installation minimizes visual distractions to maximize a listener's focus on the sounds. The primary area of visual stimulation is the interactive device. The room housing the installation is dimly lit and secluded from all other visual stimulation.

3.1 The Interactive Device

The interactive device resembles the design of a mixer as shown in Fig. 3, and is constructed using an Arduino Mega, microcontroller, with fourteen potentiometers connected to it. The housing for this interactive device is a 3D printed box. The potenti-

ometers send serial data out of the Arduino into a Max/MSP patch which was designed for this installation. The code for this device was written in the Arduino IDE.

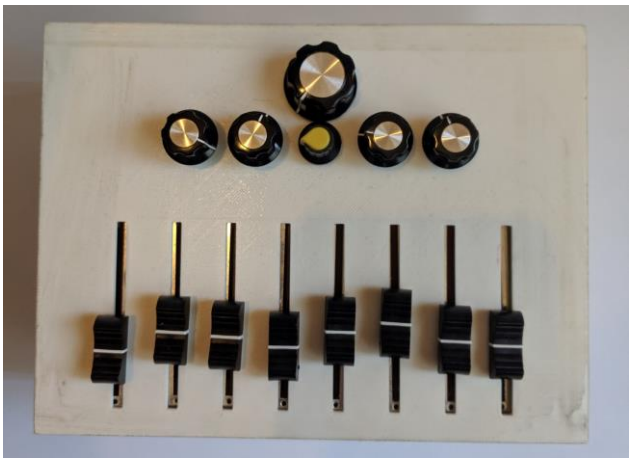


Fig. 3. Image of the interactive device

3.2 The Max Patch

The Max/MSP patch takes in the raw serial data and parses out which potentiometer is being adjusted to then augment the five distinct recorded sound objects. Sound object will have different sound traits augmented as follows:

- Sound object 1 will change its amplitude and duration
- Sound object 2 will change its amplitude and spectrum
- Sound object 3 will change its amplitude and duration
- Sound object 4 will change its speed of playback and amplitude
- Sound object 5 will change its speed of playback and spectrum

These five sound objects are then put in an Ambisonics mixing object to be heard in surround sound. All sound objects will also have their position within the mix augmented by the interactive device. Additionally, all the sound objects, except sound object 2, will stop sending out a signal after 30 seconds if a participant does not interact with the installation. These augmentations encapsulate all 5 sound traits to be observed within the installation.

3.3 The Sound Source

The five sound objects within the installation come from three distinct sources. Sound source 1 and 2 were augmented within Ableton Live 10. The sound source material was augmented to obscure the original recordings when creating the distinct sound objects for this installation.

- The first sound source is a recording of a construction truck and was edited, equalized and pitch shifted to create sound objects 1, 4, and 5.
  - Sound object 1 is supposed to sound like a vehicle of some kind but due to the pitch shifting it is hard to pinpoint the kind of vehicle it is. The length of the sound object is 30 seconds.
  - Sound object 4 is a rhythmic edited version of the recorded sound lasting 2 seconds, which was pitch shifted down in frequency and stretched. There are three distinct beats this object.
  - Sound object 5 is pitch shifted up in frequency and lasts 5 seconds. There are 9 distinct beats in this object.
- The second sound source is a recording of a person stomping around in a puddle of water on mud. The recording was edited and pitch shifted to create sound object 3.
  - Sound object 3 lasts 8 seconds and sound like an underwater recording of someone dropping an object in water.
- The third sound source is a white noise generator from Audacity and is sound object 2. No manipulation was done to this sound source. The primary purpose of this sound object was to create a noise floor that would be constantly heard while the installation was running.

4 Conclusion

I have presented *Minimal Reduction*, an interactive sound installation that focuses on the Engineer's Situation. The purpose of *Minimal Reduction* is to allow individuals to enter the Engineer's Situation to induce a state along the active listening axis. This will be most helpful for individuals who have never trained in critical listening or reduced listening. I would like to continue researching the Engineer's Situation and its efficacy in inducing a state of active listening.

5 References

1. Chion, M. "The Three Listening Modes." The Sound Studies Reader. pp. 48-53. 2012.
2. Landy, Leigh. Understanding the Art of Sound Organization. MIT Press. pp. 173. 2007.
3. "New Media Dictionary." Leonardo 34, no. 3, pp. 261-64. 2001.
4. Schaeffer, P. Treatise on Musical Objects an Essay Across Disciplines. Translated by Christine North. and John Dack. University of California Press. 2017.
5. Huberth, Madeline, J. Cecilia Wu, Yoo Hsiu Yeh, and Matthew Wright. "Evaluating the audience's perception of real-time gestural control and mapping mechanisms in electroacoustic vocal performance." Proceedings of the international conference on new interfaces for musical expression. Brisbane, Australia. pp. 206-211. 2016.

M O D U L O

Guido Kramann[1]

Brandenburg University of Applied Sciences
kramann@th-brandenburg.de

Abstract. M O D U L O is a real-time composition tool in the form of a board game and can be played on an android device. In this game there is a close connection between the musical structure and its symbolic representation on the board. The game pieces represent arithmetic operations. These are applied one after the other along a path of the shortest adjacent distances starting from a source tile representing the sequence $id(\mathbb{N}_0) = 0, 1, 2, 3, 4, ...$ The resulting altered mathematical sequences are converted into sounds. The contrahents in this two-person game, place alternately tiles on the board or move them. The goal of the game is to establish an own path by skilful moves, which consists of operations and operands as mutually different as possible and at the same time to prevent the opponent from doing so.

Keywords: algorithmic composition, arithmetic operations, realtime composition, live coding

1 Introduction

In the computer-based board game MODULO (Fig. 1), the entire development process was aimed at ensuring the closest possible connection between the state of the game and the resulting non-trivial sound.

Thus, MODULO is integrated into a series of sonified games in which an attempt is made to establish a clear connection between the course of the game and its musical implementation [5], [1]. As a special feature in comparison to the listed examples it has to be emphasized that the game structure and the game rules of MODULO were obtained directly from musical considerations. Specifically, sequences of a grammar based on arithmetic operations are generated with the help of the game moves which is called **AOG** (**A**rithmetic-**O**peration-**G**rammar).

The creation of musical structures with **AOG** is constructive. There is no harmonic analysis and no correction of the harmonic interactions. This is also not necessary because only relatively small degrees of dissonances are produced and a meaningful organization not only of sound events, but also of their harmonical relationships take place according to the generative rules [2].

Each of the opponents (white / black) in the game has a source tile. The source tile supplies the elements of the mathematical sequence $id(\mathbb{N}_0)$ at fixed time intervals ΔT: $t = 0, 1, 2, 3, 4, 5, 6, ...$ All other tiles represent mathematical

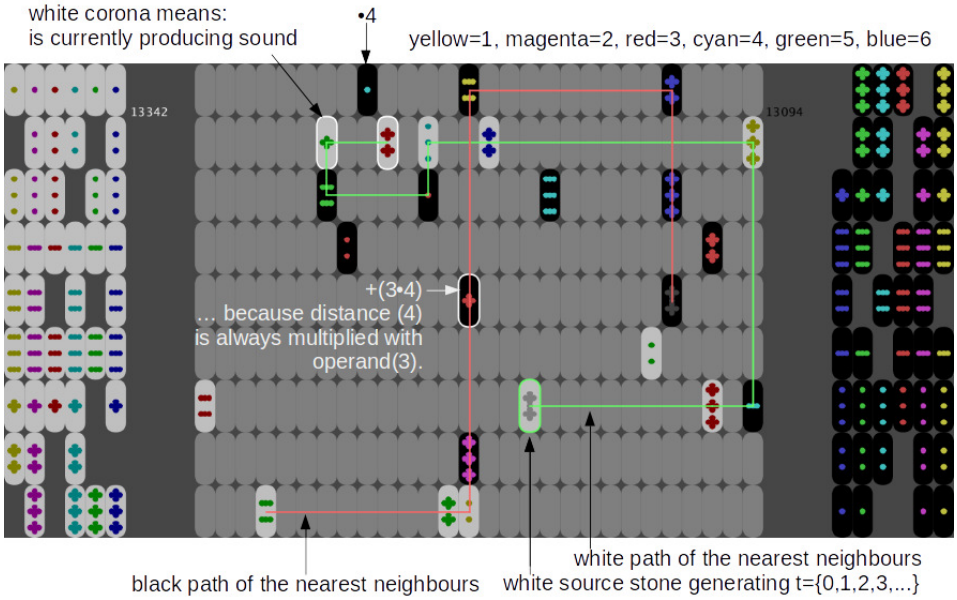


Fig. 1. View of the MODULO playing field.

operations. Starting from a source tile representing t, the arithmetic operations represented by the other tiles are applied sequentially along the path of the shortest orthogonal distances. This path is continued from tile to tile, beginning with the source tile, until one reaches a point where the condition "shortest orthogonal distance" is no longer unambiguous, or the continuation of the path would include a tile that is already part of the same previous path. The pieces included in a path can be white or black. Which player owns a path is determined by the source stone alone. These paths are automatically determined and always displayed: A green line is displayed for the path of the white player, a red line for the path of the black player.

Each tile that is not a source tile represents an arithmetic operation o_i, represented by the symbols $.,..,...,-,--,---,+,++,+++$ and an integer z_i (in the game currently: $1,2,3,4,5,6$, represented by the colors yellow, magenta, red, cyan, green, blue). The number z_i multiplied by the distance s_i to the preceding tile in the path gives the operand which is applied to the sequence t_{n-1} produced by the preceding tile, i.e..: $t_n = o_i(t_{n-1}, z_i \cdot s_i)$, see table 1 and Fig. 1.

The operators proposed here go a little beyond what is common in arithmetic. In order to understand the table, the operators $\neq, ==, \nmid, |$ should also be regarded as a type of filter that allows a number to pass when the condition meant is fulfilled. If, for example, a piece represents a division by two, the new sequence $t' = t/2$ with $t' = 0,0,1,1,2,2,3,3,4,4...$ results from t. The decimal places are omitted in all operations and values smaller than zero are set to zero. In

order to finally obtain sounds from such a sequence, each element of a sequence is interpreted as a divider d_i of a base number b – in the game b is 2520 = $2^3 \cdot 3^2 \cdot 5^1 \cdot 7^1$ – and returns the frequency $f_i = b/d_i$. f_i, however, is only considered if d_i or at least b *modulo* d_i is a true divider of b. What is more, only those frequencies are converted into sounds, which lie within a certain range. In the game it is between 55Hz and 1760Hz. This corresponds to the tones A1 to A6. This mechanism plays the role of a filter that suppresses pitches that have a too large harmonic difference to the overall structure. The sounds are represented by samples that are played by a sequencer program. If the operation of a piece produces an audible sound, it is played immediately and the piece flashes briefly (white border). The integer frequencies are mapped to the tempered tuning. Within the given limits for the frequencies and the given base number, the following scale (midi) results as a summary of all tones that can be formed: $33, 35, 37, 38, 40, 42, 44, 47, 49, 52, 54, 56, 59, 61, 63, 66, 68, 71, 75, 80, 87$.

symbol	symbol in MODULO	meaning	example
$+$	$+$	addition	$\{0,1,2,3,4\} + 3 = \{3,4,5,6,7\}$
$-$	$-$	subtraction	$\{0,1,2,3,4\} - 3 = \{0,0,0,0,1\}$
$\cdot$	$\cdot$	multiplication	$\{0,1,2,3,4\} \cdot 2 = \{0,2,4,6,8\}$
$\neq$	$++$	not equal	$\{0,1,2,3,4\} \neq 3 = \{0,1,2,0,4\}$
$==$	$--$	identity	$\{0,1,2,3,4\} == 3 = \{0,0,0,3,0\}$
$\div$	$\cdot\cdot$	division	$\{0,1,2,3,4\} \div 2 = \{0,0,1,1,2\}$
$\nmid$	$+++$	does not divide	$\{0,1,2,3,4\} \nmid 2 = \{0,1,0,3,0\}$
$\equiv$	$---$	modulo	$\{0,1,2,3,4\} \equiv 3 = \{0,1,2,0,1\}$
$\mid$	$\cdots$	true divider	$\{0,1,2,3,4\} \mid 2 = \{0,0,2,0,4\}$

Table 1: Used operators with examples.

2 Use of the game

The game is played on an Android giant tablet. The opponents sit opposite each other at a table. Thus about two times three meters of space are needed. A good system should be used for the sound. In the actual version a piano sound is used for the white pieces and a pandrum sound for the black ones. Videos of regular games as well as material of an automatically played game can be obtained here: [3]. A version only with piano sound is also available as Android app [4].

2.1 Calculation of the points gained after a move

Along the path of the player whose turn was last, the respective number of different properties in the categories color, form and number are counted for all game pieces involved except for the source tile. So you count how many different colors [1...6] there are for the pieces along the path, how many different numbers, and basic signs (each [1...3]) there are for the symbols of the involved tiles. For example the symbols (+++) and (++) have two different numbers in the appearance of their basic sign +, namely 2 and 3, whereas both have one and the same basic sign +. These three key figures (different colors, shapes and numbers) multiplied with each other result in the gain of points.

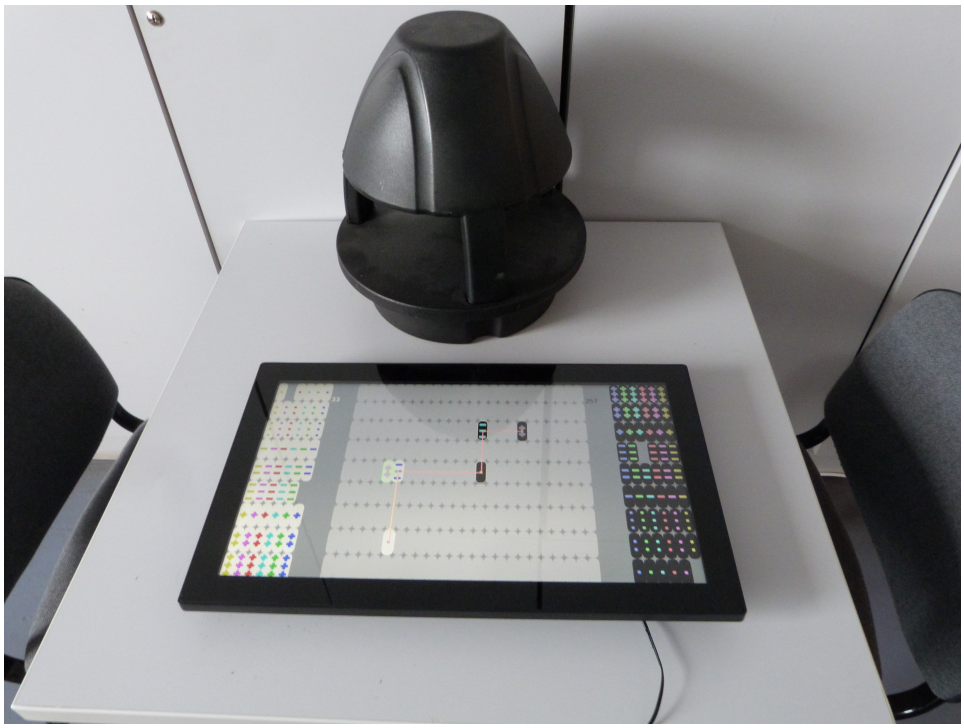


Fig. 2. Use of the game.

Fig. 3. M O D U L O on google play.

2.2 Rules of the Game

- Players can either agree to play before the start of the game until someone reaches or surpasses a certain number of points first, or set a certain fixed playing time.
- As in chess, the opponents make one move alternately.
- White begins.
- As an incentive for the opponents, the points gained are calculated and displayed after each move automatically.
- Each player has exactly one source stone in his stock.
- All other tiles in the pool represent arithmetic operations.
- A move consists of placing a tile from your own stock on the board in any free space.
- Instead, you can move any of your own tiles on the board to any free space, or put it back into the stock.

3 Strategies

You may only move tiles of your own colour, but the paths are formed taking into account all tiles lying on the playing field. This way, enemy structures can be used or disturbed.

Not all actions on the playing field lead to immediate changes in the musical structure, but they prepare it in so far as a later action can result in a path which then includes the previously musically inactive elements.

Metamorphosis. Typically, an existing path that represents the successive mathematical operations is extended by one element with another move. This causes the existing related mathematical sequences to be extended by another relatively similar one. In terms of sound, this means that another voice appears that varies the existing one (Fig. 4).

Context switch. If, during the course of a path, an element is suddenly moved or added to a location and another neighbor appears as the next element, the path takes a different course from this location after this move. In this way, several sequences are typically exchanged at the same time. This type of change takes place rather later in the course of the game, this is when greater changes are expected from the point of view of musical dramaturgy either.

Hostile takeover. If, during the course of a path, an element is suddenly moved or added to a location and another neighbor appears as the next element, the path takes a different course from this location after this move. In this way, several sequences are typically exchanged at the same time. This type of change takes place rather later in the course of the game, i.e. when greater changes are expected from the point of view of musical dramaturgy (Fig. 5).

Blockade, sudden silence. For both opponents the path is constantly formed, which always leads from the own source piece to the next neighbour, until this rule can no longer be applied unambiguously, or a piece already integrated into the path has to be connected. If a player causes such a ambiguity in the opponent's path by placing a tile in the neighborhood of a tile involved

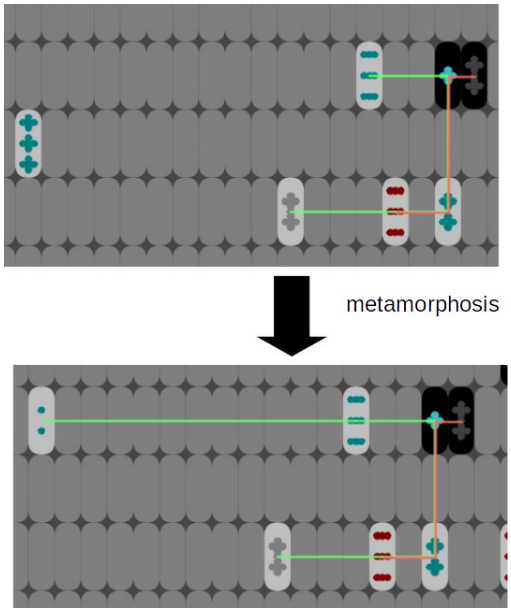


Fig. 4. Addition of an operation towards an existing path (metamorphosis).

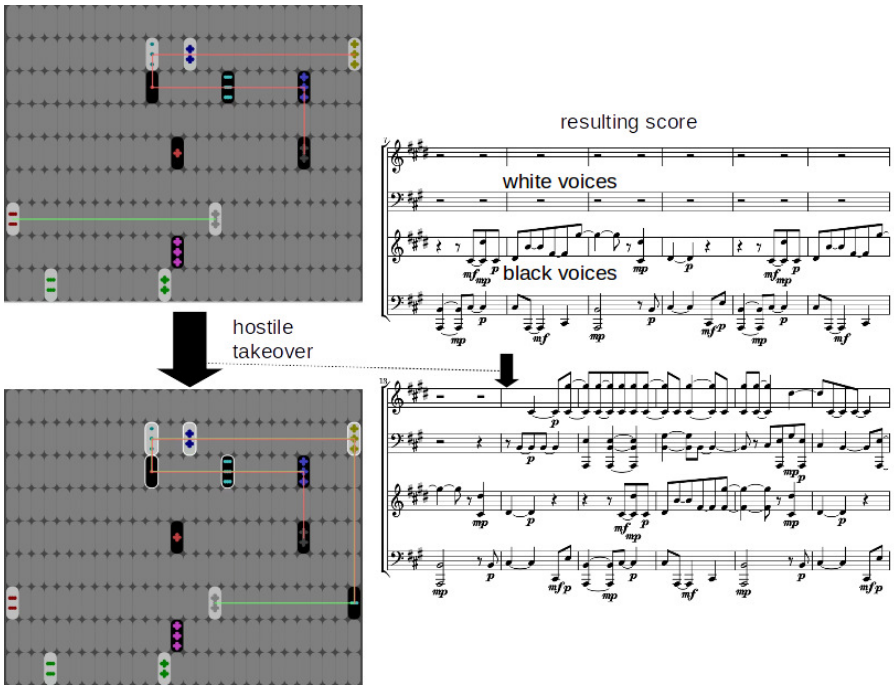


Fig. 5. Switching path by adding element close to source tile (hostile takeover).

in the opponent's path in such a way that it lies at the same distance as the nearest neighbor there, the complete following path disappears immediately and in extreme cases sudden silence occurs.

4 Discussion and Further Work

On the one hand it is enough to master the few simple rules to play MODULO.

However, the basic idea of MODULO is also to create the musical events through the own way of playing perhaps not exactly fully conscious, but at least with time to get a feeling for how actions on the game board affect the sound level.

This is promoted in particular by the fact that the way in which the modified sequences of numbers and the resulting sound events can be determined from a path is clearly and quite transparently predetermined by the generative grammar on which the game MODULO is based.

In concrete terms, it is not so difficult to carry out the unfolding process from the symbolic representation on the playing field to the music score behind it with pen and paper for smaller paths.

If, on the other hand, the resulting musical phrases and harmonies were to be interpreted in more detail using various instrumental playing techniques, for example by using physical modeling to generate sound, this could also be experienced in a more in-depth way as the musical representation of the structural interrelationships on the playing field by the players, thus promoting intuitive understanding of the game in another way.

References

1. Hamilton R.: Musical Sonification of Avatar Physiologies, Virtual Flight and Gesture. In: Sound, Music, and Motion (CMMR 2013), pp. 517–532, Springer, Heidelberg (2014)
2. Kramann, G.: Generative Grammar Based on Arithmetic Operations for Realtime Composition. In: CMMR 2019 (2019)
3. Kramann, G.: M O D U L O infos `http://www.kramann.info/cmmr2019b` (2019)
4. Kramann, G.: M O D U L O app `https://play.google.com/store/apps/details?id=kramann.info.MODULO&gl=DE` (2019)
5. Sinclair S., Cahen R., Tanant J., Gena P.: New Atlantis: Audio Experimentation in a Shared Online World. In: Bridging People and Sound (CMMR 2016), pp. 229–246, Springer, Heidelberg (2017)

A Real-time Synthesizer of Naturalistic Congruent Audio-Haptic Textures

Khoubeib Kanzari[1,2,3], Corentin Bernard[1,2,3], Jocelyn Monnoyer[2,3], Sebastien Denjean[2], Michaël Wiertlewski[3,4], and Sølvi Ystad[1]

[1] Aix Marseille Univ, CNRS, PRISM, Marseille, France
{kanzari;bernard;ystad}@prism.cnrs.fr
[2] PSA Groupe
{jocelyn.monnoyer;sebastien.denjean}@mpsa.com
[3] Aix Marseille Univ, CNRS, ISM, Marseille, France
[4] Delft University of Technology, The Netherlands
m.wiertlewski@tudelft.nl

Abstract. This demo paper presents a multi-modal device able to generate real-time audio-haptic signal as response to the users' motion and produce naturalistic sensation. The device consists in a touch screen with haptic feedback based on ultrasonic friction modulation and a sound synthesizer. The device will help investigate audio-haptic interaction. In particular the system is built to allow for an exploration of different strategy of mapping audio and haptic signal to explore the limits of congruence. Such interactions could be the key to more informative and user-friendly touchscreens for Human-Machine-Interfaces.

1 Introduction

The mental representation of the environment is shaped by a multitude of sensory inputs integrated across space, time and sensory modalities. The perception of a same signal is reinforced by being congruently detected via multiple modalities such as touch, audition and vision. In this context, several studies showed that the cross-modal integration of haptic and visual feedback is achieved in an optimal way following a maximum likelihood principle [1]. However, audio-haptic interactions are less understood. In some instance, audition influences touch while exploring a texture. For instance, the *Parchment-skin Illusion* describes an modification of the tactile perception of roughness due to an attenuation of the high-frequencies of the sound produced by the interaction with solid objects [2]. This effect has been rigorously demonstrated on abrasive papers [3]. The judgement of tactile roughness is also altered when listening to synthetic sounds instead of recorded, touch-produced sounds [4]. For the effect to reach his full potential auditory and tactile stimuli must be congruent both in intensity [5] and spatially [6]. Audio and haptic feedback have already been associated to guide the user gesture on an interface [7] [8], but the expected performance improvement by combining this modalities is not evident. Investigating correlations between texture parameters and sound [9], it was demonstrated that

texture roughness is better perceptively match to the sound intensity than to the pitch.

This paper describes a device able to produce congruent audio-haptic signals in real-time as a function of the user motion, using a combination of a surface-haptic device and a sound synthesizer. Investigations of the audio-haptic interactions by simulating a wide range of tactile stimuli and realistic sounds are possible. Better understanding of cross-modal audio-haptic interactions could be the key to more informative and user-friendly touchscreens for Human-Machine-Interfaces.

2 Apparatus

2.1 Audio synthesizer of interaction sounds

The sound synthesizer is based on perceptually relevant sound morphologies associated with the recognition of objects (structural invariants) and actions (transformational invariants) in line with the ecological approach to perception proposed by Gibson [10]. The identification of invariant sound structures led to an *Action-Object Paradigm* for sound synthesis in which sounds are described as the consequence of an action (like knocking, rubbing or scratching) on an object, defined by its shape (width, thickness, curvature...), and material (metal, wood, glass, ...). The action is modeled by a low-pass filtered noise simulating successive impacts and the object by a resonator, implemented through a resonant filter bank. A perceptual mapping enables continuous transitions between the different categories of actions and objects [11].

2.2 Ultrasonic friction modulation technology description

The haptic surface device uses ultrasonic levitation to modulate the friction between a glass plate and the user's finger. The glass plate vibrates at a resonant mode 35 kHz over amplitude of 3 microns. These vibrations are not directly perceptible by the skin, but they cause the finger to slightly levitate from the plate and dramatically reduces the friction between the finger and the screen [12]. Modulating the amplitude of the ultrasonic vibration allows for a controlled modulation of the friction force applied to the finger. Patterning friction as a function of position and velocity of the finger can thereby create the illusion of touching shapes and textures [13].

2.3 System architecture

The global operation of the audio-haptic synthesizer is described in Fig. 1. Haptic rendering is acheived by a simple texture generator. The texture is encoded by three parameters : a sine wave described with fixed spatial frequency f_s and amplitude, A and an entropy value h referring to the proportion of noise. The resulting 1D friction map μ_{map} is thus calculated on the micro-controller and updated each time the parameters f_s, A, h are modified on the user interface. μ_{map}

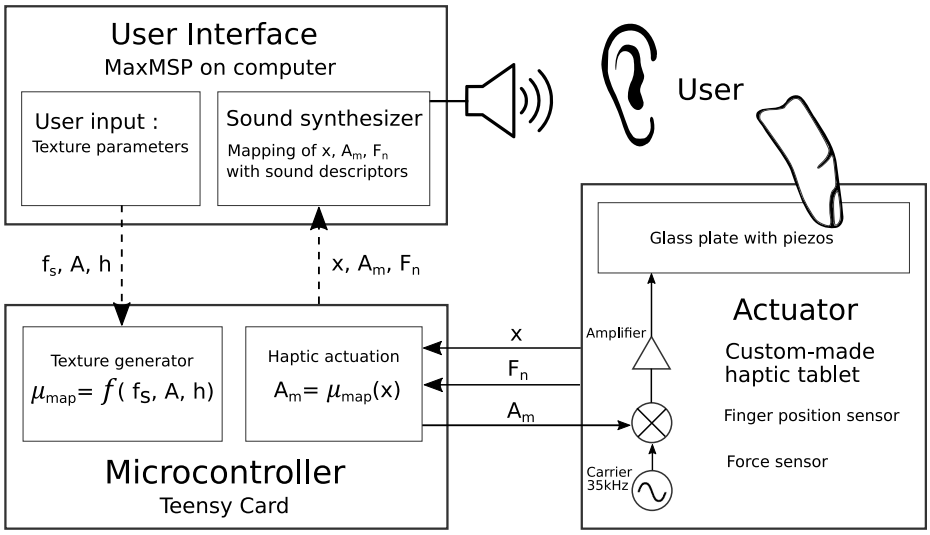


Fig. 1: Description of the audio-haptic synthesizer operation. Analog communications are printed in continuous lines and digital communications in dashed lines.

describes the friction level according to the finger position x. While running, the haptic signal generation is performed by a real-time loop that first acquires the user's position x and outputs the modulation amplitude A_m stored in the friction map. A_m is then converted into an analog signal modulating a high frequency carrier wave, and therefore the frictional behavior of the plate. Simultaneously, the finger position x, the modulation amplitude A_m and the normal force applied by the finger on the plate F_n are sent by the microcontroller to the Max/MSP sound synthesizer through a serial port. These parameters are used to control the sound synthesizer in real-time.

3 Audio-haptics signal mappings

A prerequisite for the study on audio-haptic interaction is to recreate sounds that are perceived as congruent with the explored haptic texture. In this perspective, we investigated different mapping strategies between the haptic device and the synthesized sound. Previous studies [14] have shown that a finger movement on a surface can be simulated by mapping the finger velocity to the cut-off frequency of a low-pass filter. In this study we propose to proceed in the same way by varying the attributes of a band pass filter (f_0 , B and G) according to the contact properties of the haptic surface ($v = dx/dt$, F_n and A_m). We tried several possible mappings, but opted for the one in Fig. 2 through non-formal tests. We chose to apply this filter on white noise in order to demonstrate the

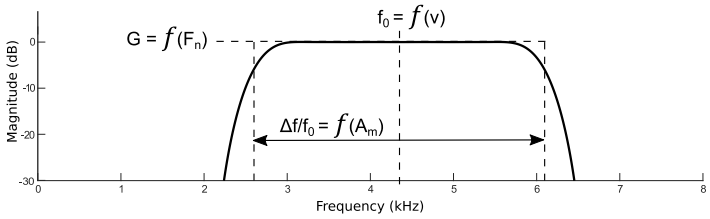


Fig. 2: Mapping between properties of the haptic surface and the bandpass filter attributes. The central frequency f_0, the relative bandwidth $\Delta f/f_0$ and the gain G depend respectively to the finger velocity $v = dx/dt$, the current friction level A_m and the normal force applied by the finger F_n.

audio-haptic congruence on a neutral sound before generalizing this principle to more complex sounds textures.

The finger velocity, calculated as the time derivative of the position, is mapped to the central frequency f_0 of the bandpass filter. This central frequency increases linearly as the finger speeds up. Moreover, the current friction level A_m determines the relative bandwidth $\Delta f/f_0$. They are inversely proportional according to a logarithmic profile. Indeed, a high value of A_m corresponds to a minimum bandwidth. On the other hand, the normal force F_n exerted by the finger on the glass plate is collected to determine the gain G of the filter in a linear way.

From a perceptual point of view, with this synthesis method, sliding the finger with a higher velocity results in a brighter sound. Moreover, while sliding the finger on the ultrasonic surface, we perceive the rendered haptic effect as a periodic succession of low and high friction areas, with their intensity, frequency and regularity depending on the parameters chosen by the user. With the proposed mapping, slippery regions of the surface area are associated to a sound with a richer frequency content than the sticky area. It results in hearing pulses when the finger crosses theses regions. Besides, an increase of the normal force applied by the user leads to a higher sound intensity.

We noticed through non-formal tests that an additional mapping, in which the velocity also influences the relative bandwidth, is also perceptually interesting and strengthens the congruence between the sound and the finger movement.

4 Conclusion

In this study, we aim to evaluate audio-haptic perception by combining an ultrasonic friction modulation device with a sound synthesizer that simulates environmental sounds. Both signals are simultaneously generated in real time and the parameters from the haptic device directly control the properties of the bandpass filter applied to the synthesized sound. During the demo session several mapping strategies and sound textures will be proposed and discussed in the light of previous and future perceptual evaluations of the device.

References

1. Ernst, M.O., Banks, M.S.: Humans integrate visual and haptic information in a statistically optimal fashion. Nature **415**(6870) (January 2002) 429–433
2. Jousmäki, V., Hari, R.: Parchment-skin illusion: sound-biased touch. Current Biology (1998)
3. Guest, S., Catmur, C., Lloyd, D., Spence, C.: Audiotactile interactions in roughness perception. Experimental Brain Research (September 2002)
4. Jir, G.: Effects of Auditory Feedback on Tactile Roughness Perception. 13
5. : Effects of sounds on tactile roughness depend on the congruency between modalities. In: World Haptics 2009, Salt Lake City, UT, USA
6. Gyoba, J., Suzuki, Y.: Effects of Sound on the Tactile Perception of Roughness in Peri-Head Space. Seeing and Perceiving (2011)
7. Rocchesso, D., Delle Monache, S., Papetti, S.: Multisensory texture exploration at the tip of the pen. International Journal of Human-Computer Studies **85** (January 2016) 47–56
8. Rocchesso, D., Papetti, S.: Path Following in Non-Visual Conditions. IEEE Transactions on Haptics (2018) 1–1
9. Peeva, D., Baird, B., Izmirli, O., Blevins, D.: Haptic and sound correlations: pitch, loudness and texture. In: Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004., London, England, IEEE (2004) 659–664
10. Gibson, J.J.: The Ecological Approach to Visual Perception. 347
11. Conan, S., Aramaki, M., Kronland-Martinet, R., Thoret, E., Ystad, S.: Perceptual differences between sounds produced by different continuous interactions. (2012)
12. Wiertlewski, M., Friesen, R.F., Colgate, J.E.: Partial squeeze film levitation modulates fingertip friction. Proceedings of the National Academy of Sciences (2016)
13. Bernard, C., Monnoyer, J., Wiertlewski, M.: Harmonious Textures: The Perceptual Dimensions of Synthetic Sinusoidal Gratings. In Prattichizzo, D., Shinoda, H., Tan, H.Z., Ruffaldi, E., Frisoli, A., eds.: Haptics: Science, Technology, and Applications. Volume 10894. Springer International Publishing, Cham (2018)
14. Thoret, E., Aramaki, M., Bringoux, L., Ystad, S., Kronland-Martinet, R.: Seeing Circles and Drawing Ellipses: When Sound Biases Reproduction of Visual Motion. PLOS ONE **11**(4) (April 2016) e0154475

CompoVOX 2: GENERATING MELODIES AND SOUNDPRINTS FROM VOICE IN REAL TIME

Daniel MOLINA[1,2] Antonio JURADO-NAVAS[1] Isabel BARBANCHO[1],

[1] University of Malaga,
[2] University Jean-Monnet,
danielmolinavillota@gmail.com

Abstract. This project involves the development of an interactive application for musical sequence generation from human voice. There are used different characteristics of voice, as vowel sounds, central frequencies and level. Differences sequences of programming are developed in MAX MSP to parallelly make different layers of sounds related closely to properties of voice signal before stated. This application also allows to visualize a graphic interface that changes with the sounds produced by this software. At the end, we obtain an automatic and tonal musical composition and a sound print of human voice.

Keywords: Sonification, treatment of sound signal, voice, automatic composition in real time, tonal music, treatment of voice.

1 Introduction

Actually, many digital instruments and programs are focused in creating music and sound predominantly from touching or moving body parts but also from digital interface graphic objects mostly in video games, where is very suitable use automatic sound to relate distance and sound pressure level. Usually this kind of applications generates or employs graphics that helps to summon speaker/viewer to play again. Then create an interactive application that doesn't use movement or buttons like traditional systems was the goal of this project. Voice as a generator of musical sequences, can be interesting both for composers and amateur users, since they can explore aspects of the voice for sound generation and also to explore a sound print of voice. In human communication, voice is the most important tool and it's also our major musical instrument, therefore its use can be quite interactive. CompoVOX2 has been developed in MAX MSP using data processing, visual, and sound effects tools.

2 Talking About Interactivity

Grace to advances processing capacity, today we can use software that back in the time was not possible. It was necessary an extensive quantity of informatics resources

that today are in just one computer of high performance, then signal treatment and projects involving digital art and engineering is so much achievable today. From an artistic theorist aspect, this project tries to apply the concept of synesthesia and non-synesthesia, usually all musical instruments generate sound from moving of our fingers, today there are also projects that try to generate sound from moving arms, body, etc. Also, for graphics we have seen several projects to generate graphics from body moving and fingers moving, mostly in video games. Then all this application are based in synesthetic process where a mechanical signal is transformed in electric signal (or digital) and generates images or sound, but what happens if we change the principal source of this kind of creation, we can generate sound using another sound, and for human, the more important sound is the voice.

Here we will mention some projects that have been important for us to think in this project. Opto-isolator [1], is an integrative project who induces the viewer to a high interaction, it's proposed that it would be the work itself who observes the viewer. Re: MARK, is another project that uses voice to create image from voice's analysis (identification of phonemes), and the movement of the participants to produce real-time animations. The importance of this project is primary, that there is developed an interface who allows viewer to play actively. Additionally, projects like WIP [4] are visually interesting, taking curve and the amplitude of sound to allow in real time generate multiple visual combinations, projection of the geometric shapes and appearance modification.

Now we see easily how this project begin with a proposition opposite to the previous works noted, but also employing an important fact in all of them, a graphic interface who inspire to be played. This work is achievable using real time synesthetic and no synesthetic procedures to generate sound and image, and at same time creating a tonal music sequence from voice and taking to generate visual forms related to the sound performed on the screen shown in the installation. A first step of this project Called CompoVOX was developed under these principles [5].

Now a second application has been developed based on a deeper treatment of signal and also employing new techniques to generate music and also a sound print of human voice.

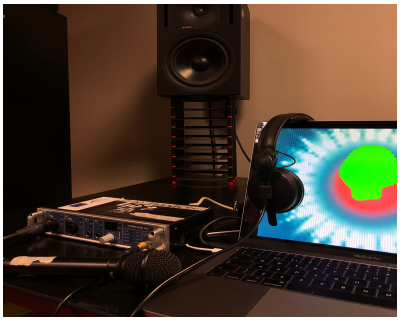


Fig. 1. Compo VOX

3 Sound Synthesis, Graphics and Sound print of Voice

At first, we have taken different sound sources and store it so we could do different experiments of filtering and using voice to generate music. It has been employed a dynamic microphone. This microphone allows to isolate the environment, and, in this way, it is possible to focus only on the capture of voice, of course, an audio card is used, and also a software to record sound, in this case we have use also MAX MSP. The program is based on the first version of CompoVOX and is responsible for taking the audio signal and through different stages of treatment, obtain a signal that can be translated into MIDI values and for so in a tonal musical sequence.

To mention sound generating, we have to talk about different types of voices like woman, man, children, then we have employed 3 different filters based on central frequency of talking voice, in this way, the region of interest for this project is filtered employing a digital FFT (Fast Fourier Transform) Filter that cuts the signal in different intervals, for example for male voice, after recognize central frequency we will use just spectrum ranges between 50 and 600 Hz. Once this process has been carried out different Fourier filters are performed in parallel, which focus on taking just the signal of certain regions of the spectrum. To define such regions of interest, several tests have been made by differentiating the levels reached in each region, level is responsible for activating different areas of the musical scale, usually the low notes activate lower sounds and the high notes will activate higher sounds. Our final generator of music will be composed of experimental techniques and also more advanced techniques that we will treat later in this paper.

Signals come directly from the voice taken from a microphone, then level signal will vary quickly, therefore, to smooth that fast changes, it's necessary employ an averaging function, this averaging avoids sudden variations, and eliminates possible intrusions of sound from the environment. A signal that varies in slower way facilitates the control in real time of musical parameters that must be audible for the user.

We talked before about two steps of filtering to identify voice and also to use it for music generation, then it is very important mention that entire system is controlled by the one clock, but each synthesis stage is activated only when an appropriate level range is reached in the indicated region of the spectrum and in that case. There will be a mapping of MIDI notes that passes through a tonal filter, it makes the system more regular and keep it in just one key.

Different parts of the signal's spectrum are used to control notes and its parameters like attack and the temporal length. On the other hand, another process of sound synthesis is done by using the central frequency of the voice and its variations. Those parameters help to generate a sequence of notes that is passed through the same tonal filter before noted, and subsequently the frequency value serves as a tone control parameter of another synthesizer.

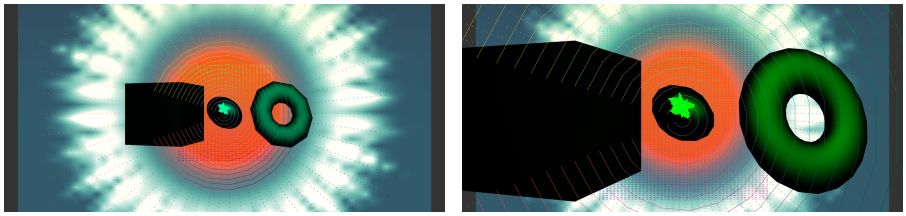


Fig. 2. Objects moving in graphic interface

Now we will talk shortly of graphic interface. A background treated by MAX MSP is responsible for reproducing a space environment and comes from noise of original signal. There is a conical gang that covers the space and that varies in sharpness and size proportionally to the central frequency of the voice. A flat circular object located in the center takes different forms correspondingly to the waveform of the voice. And several objects change shape according to the level in a region of low frequencies of the spectrum. The whole system moves proportionally to the level of sound, that is, the louder the voice is, the closer the objects are, and the weaker they are, the farther they go. The frequency is also responsible for changing the color of the objects.

How we saw before we have established a method for automatic music generation from voice in real time, an also a basic mechanism of synesthetic interactivity, that means changing graphics in size, distance and color. Now we are interested in creating a major interaction, for that reason we have employed a method to react to four simple instructions, to go octaves up or down or to do lower or faster musical sequences, employing words like up, down, low and fast [6].

We use also a formant analysis for talking voice (it does not apply for singing voice where we prefer to identify active harmonics for the sonification process), in first place we will do vowel recognizing considering energy and no passing zero-axe, vowel signal is kept in a window of signal. After we use FFT to find first and second formant frequencies of one person. These formant frequencies will be used to generate what we call sound print of voice and sound generated from this voice. We know then about musical aspects of timber of each person. [7,8]

Finally, for graphics we employ also movement up and down for objects corresponding to higher or lower tune of musical sequences creates before mentioned.

4 Performing This Demo

This demo is straightforward to employ. Thus the user has just to talk by microphone. In real time, his/her voice will be filtered. Voice will also be used as a mean of controlling and generating patterns that will be reproduced by the sequencer. This filtering and sequence will generate a synesthetic visual analogue that will show

details of the participant's voice spectrum. Additionally, the participant will be able to play with a rhythmic base that changes tonality with central frequency of talking or singing voice. Music will be bounded to graphics that changes color and position depending on tune.

5 Conclusions

Sonification of voice signal for the generation of tonal musical sequences accompanied by a graphic interface related directly to voice is highly interactive and allows exploring aspects of sound generation that are not traditionally used in music.

CompoVox2 also helps to create individual sound prints of voice.

References

1. G. Levin, "OptoIsolator". Project at M.I.T Massachusetes technological Institute. 2007.
2. Z. Lieberman, G. Levin, "Re: MARK". Project for FutureLab of Ars Electronics (Siemens). 2002.
3. N. Nordmann, J.R.Valentin, "Variationen für Teno-ri-on". Project presented at Kunstvolkslauf Hanno-ver, Zinnober in 2013.
4. B. Guesnon, "W.I.P Work.In.Processing". Project by himself begun at 2013.
5. D. Molina, I. Barbancho, A. Jurado-Navas. Real Time Sonification of Voice. Proceedings 16th Sound and Music Computing Conference 2019.
6. J-F. Charles. A Tutorial on Spectral Sound Processing Using Max/MSP and Jitter Computer Music Journal Vol. 32, No. 3, Synthesis, Spatialization, Transcription, Transformation (Fall, 2008)
7. Rabiner, Lawrence and Biing-Hwang Juang. Fundamentals of Speech Recognition. pp1-65. Prentice-Hall, Inc. c.1993.
8. K. D. Martin. Sound-Source Recognition: A Theory and Computational Model.

Coretet: a 21st Century Virtual Interface for Musical Expression

Rob Hamilton[1]

Rensselaer Polytechnic Institute
hamilr4@rpi.edu

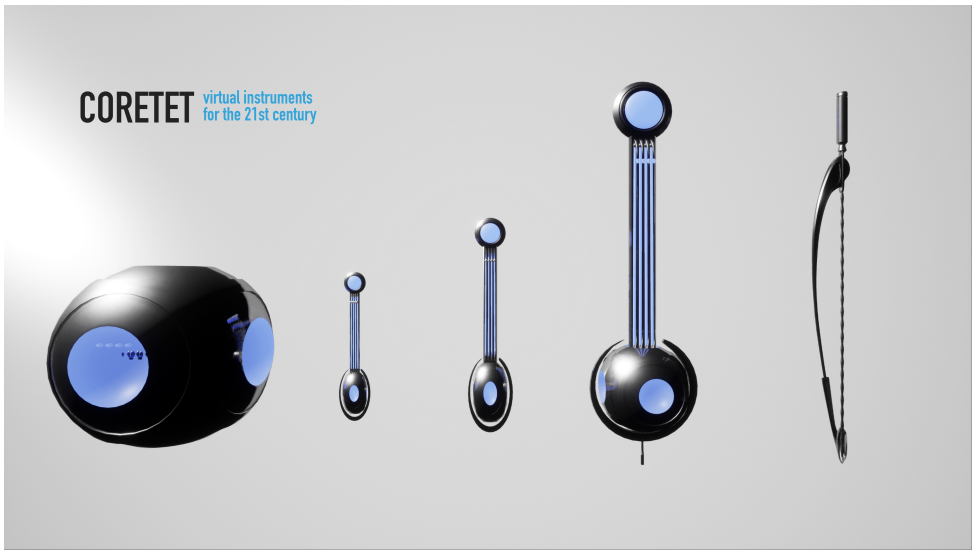


Fig. 1. (*from left to right*) Coretet instruments: the orb, violin, viola, cello, bow

Abstract. Coretet is a virtual reality musical instrument that explores the translation of performance gesture and mechanic from traditional bowed string instruments into an inherently non-physical implementation. Built using the Unreal Engine 4 and Pure Data, Coretet offers musicians a flexible and articulate musical instrument to play as well as a networked performance environment capable of supporting and presenting a traditional four-member string quartet. This paper discusses the technical implementation of Coretet and explores the musical and performative possibilities enabled through the translation of physical instrument design into virtual reality as realized through the composition and performance of the string quartet *Trois Machins de la Grâce Aimante*.

Keywords: Virtual Reality, VIME, VMI, instrument

1 Introduction

With renewed interest in virtual reality (VR) hardware and software systems in recent times by major companies such as Facebook, Google, HTC and Valve, consumers and artists alike have multiple low cost paths forward to investigate VR systems as a component in their personal and artistic practices. Commodity control devices for VR systems such as the Oculus Touch allow three-dimensional tracking of heads and hands with sufficiently low latency to allow for real-time gestural control of musical systems and instruments [19]. Laptops and graphics cards have likewise reached price points and speeds capable of delivering consistently high framerates for head-mounted displays (HMD) in packages that are portable and easily replaced.

These factors contributed to the design and development of Coretet, a real-time VR instrument modeling basic bow and string interactions and performance practices idiomatic to stringed instruments such as violins, violas, cellos and doublebasses. Commissioned by the GAPPP project [13] at the IEM in Graz, Austria, Coretet - and the first composition for Coretet *Trois Machins de la Grâce Aimante* for virtual reality string quartet - showcase commodity technologies such as Oculus Rift HMDs and the Unreal Engine 4 augmented with Open Sound Control (OSC) [23] and a Pure Data (PD) [18] audio engine driving a physical model of a bowed string from the Synthesis Toolkit (STK) [3].

2 Prior Art

Software instruments derived from traditional instrumental performance practice and the act of composing musical works for such instruments have long been a staple of computer music research and artistic practice. Computer-mediated and augmented stringed instruments have been explored [2, 15, 16] that blend digital control systems with traditional musical performance practices and leverage expert performers' learned instrumental mastery. Commercial gaming applications such as the Guitar Hero and Rock Band [10] game franchises introduced score-driven and rhythm reliant performance practices for pop and rock music, while mobile apps like Magic Fiddle [22] allowed performers to interact with a rendered violin neck and strings using finger gesture to activate a bowed or plucked string excitation on an iPad tablet interface. And immersive musical environments for musical composition and performance combining 3D graphics and procedural sound have been explored, both by extending existing software projects using Open Sound Control [8] or by creating entirely new projects using game development platforms like the Unreal Engine [9].

Virtual interfaces for musical expression have been explored using Cave-like 3D displays [14], haptic interactions [11] and by leveraging game-engines as mediating layers for user interaction [5, 21]. And as part of a 1993 exhibit on virtual reality at the Guggenheim Museum in New York, Thomas Dolby presented *The Virtual String Quartet*, a pre-recorded and non-real time animated performance of Mozart's String Quartet no. 18 in A Major viewable through HMDs and spatialized by tracking audience members' positions in a gallery space [4].

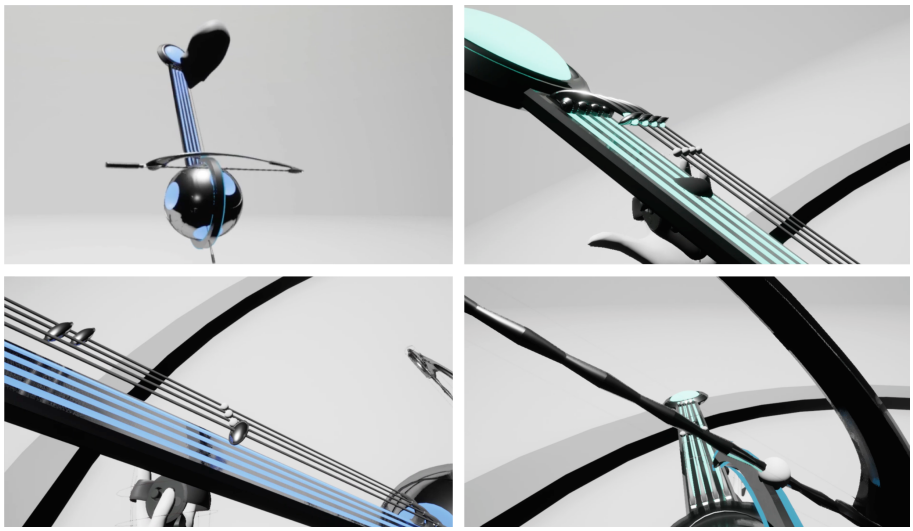


Fig. 2. (*clockwise from top left*) Figure 2a: Coretet client's head, bow and cello viewed on the server; Figure 2b: client view of left-hand at the top of the instrument neck; Figure 2c: the bow contacting the bowing bar; Figure 2d: left hand selecting a pitch on the lowest string.

3 Design and Development

With deliberate reference to traditional stringed instrument performance practices, Coretet was designed as a futuristic '21st Century' implementation of the core gestural and interaction modalities that generate musical sound in the violin, viola, cello and doublebass. Coretet's primary design goals focus on the creation of virtual musical instruments intended for expert users, defined here as trained musicians with substantial history playing and tacit knowledge of traditional bowed string instruments. And through the leveraging of trained performers' tacit learned knowledge of traditional stringed instrument performance practices, Coretet's design aims to allow performers to achieve a high level of skill on the instruments eventually leading towards virtuosity.

3.1 Control and Interface

By coupling the real-time fluid interactivity offered by the Unreal Engine 4 game engine with procedurally driven STK physical string models in Pure Data, Coretet offers musicians a flexible and articulate musical instrument capable of supporting improvisation as well as notated solo or ensemble performance. Fundamentally Coretet is a single instrument which can be shaped and scaled by performers into different configurations (see Figure 1). Parameters such as neck length, body size, and number of strings are manipulated to recreate traditional stringed instruments such as violin, viola, cello or doublebass or to create new

and physically impossible instruments. For ease of use during performance parameter presets for violin, viola, cello and doublebass can be chosen and recalled instantly as can an experimental spherical percussive instrument configuration known as the Orb.

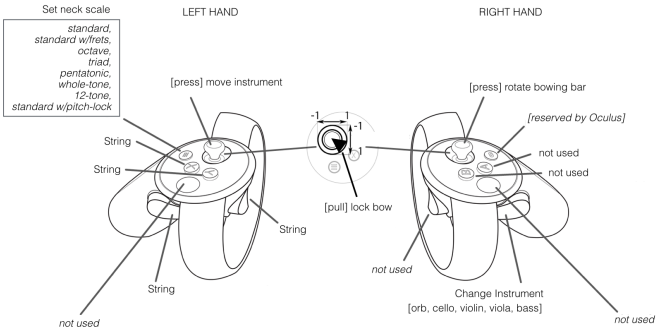


Fig. 3. Left and right hand Oculus Touch controller mappings for the Coretet instruments.

Performers use a virtual bow (see Figure 2a) modeled after a traditional violin bow which activates the bowed string physical model when it comes into contact with a specific bowing bar on the instrument. Figure 2c shows a blue outline around the bowing bar indicating a collision between bow and bar, and a tracking marker indicating the collision is represented as a white sphere on the bow. Bow pressure is controlled by calculating position along the bowing bar with one end representative of a high level of bow pressure and the other end representative of a low level of bow pressure. Similarly, bow speed is calculated by windowing bow positioning deltas over a short time frame.

By pressing buttons on the Oculus Touch's left hand controller, performers choose which string will be activated (see Figures 2d and 3). By moving their left hand along the instrument's neck and pressing each string's associated button, performers change the pitch of the current sounding note. String positions activated by button presses are marked in real-time by dark-grey oval markers (see Figure 2b and 2d). For *Trois Machins de la Grâce Aimante* each string of Coretet is tuned to the same fundamental frequency as the corresponding string on the violin, viola and cello in concert A 440 Hz tuning.

3.2 Networked Environment

Coretet leverages the Unreal Engine's native network layer to create a networked virtual performance environment capable of supporting and presenting a tradi-

tional four-member string quartet to performers through head-mounted displays and to audiences through an auxiliary screen or projector. In a concert performance such as is utilised for *Trois Machins de la Grâce Aimante* this game server hosts each Coretet client instance (representing each performer) connecting across a local ethernet network. Performers in Coretet see each others' head, bow and instrument in real-time within the virtual concert space, allowing for the use of communicative visual gesture both of the head and of the instrument and bow. In live concert situations, a view into the networked virtual space is presented to audiences from the game server. In a manner similar to e-Sports broadcasts of networked games, a series of virtual cameras on the server are projected in 2D for viewing by audiences seated in traditional concert halls.

3.3 Procedural Audio

The bowed string sounds used in Coretet for *Trois Machins de la Grâce Aimante* are generated in real-time using Pure Data. String length and body scale values of the current instrument configuration are tracked in UE4 and sent to a single running PD server via OSC where they are used to procedurally generate each instrument's sound affecting it's frequency range and timbral identity. Within Coretet, OSC is implemented using a plugin for Unreal's Blueprint workflow programming language. Currently, a single Pure Data instance receives OSC data from all connected Coretet instruments and drives four mono output channels with one speaker assigned to each performer.

Within PD the *bowed~* physical string model from the STK as found in the PD port of the PerCoLate package receives parameters including bow pressure, string position, velocity, and string length. The sound of the physical model is augmented with low-volume sine oscillators, and fed through a series of effects including gain staging, reverberation and a compression/expansion process. Additionally, percussive sounds in Coretet as triggered through virtual hand collisions with the sphere of the Orb configuration are generated using the STK *agogo~* model.

3.4 Parameter Mappings and Modes

To map performative gesture and interactions between bow and instrument to salient parameters of the Coretet audio system, a series of key parameter mappings are used.

Bow velocities. The velocity of the moving bow is tracked at two distinct locations: the frog (near the performer's hand) and the far tip of the bow. At this time only the windowed frog velocity is used to drive each instrument's gain level. Frog velocity is additionally used as a factor in scaling the current detected level of bow pressure.

Bow and bowing bar collision locations. When bow collisions with the bowing bar are detected, the position of bow collision along the curved surface of the bowing bar is used to calculate bow pressure to be applied to the string model. Due to the current practice of attaching the bow to the player's VR hand, a point of collision cannot be tracked in UE4 using standard techniques. Instead, along the bowing bar static mesh are embedded 32 tracking nodes. Similarly 12 nodes are embedded across the length of the bow. The closest distance calculated between bow nodes and bowing bar nodes is used to approximate the point of contact.

String length. When different instruments are selected the scale and length of the instrument neck changes. The string length of any selected string is used alongside the desired tuning of each string to calculate the current sounding note.

Finger Position. Any selected finger position on the neck is used alongside the current string length to calculate the current sounding pitch of the string. When quantized scales or alternate modes are selected on the instrument, the exact finger position is fed into the rounding algorithm for the desired scale or mode.

Neck scale/mode. When selected, finger positions along the neck of the Coretet instrument can be quantized to a variety of modes and scales. To denote each selected mode or scale fret markings similar to those found on a viola da gamba or guitar are made visible along the instrument's neck. These modes and scales include:

- Octave: the neck is divided into two regions.
- Triad: major triad built on a string's root pitch.
- Pentatonic: a five note scale.
- Whole-tone: a six note whole tone scale.
- Chromatic: a single octave chromatic scale.
- Quantized: the full range of the instrument with pitches quantized to a chromatic diatonic scale.
- Free: the full continuous range of the instrument without quantization.

Hand collisions. For the Orb configuration of Coretet, hand collisions with the instrument's body will trigger a pitched percussion note using the agogo model. For *Trois Machins de la Grâce Aimante* triggered notes are selected from a composed set of pitch classes. Hand velocity controls the gain of the signal.

3.5 Notation and Scoring

Within a virtual reality environment where users wear head-mounted displays, performers are unable to view notated scores in a traditional manner. For the first

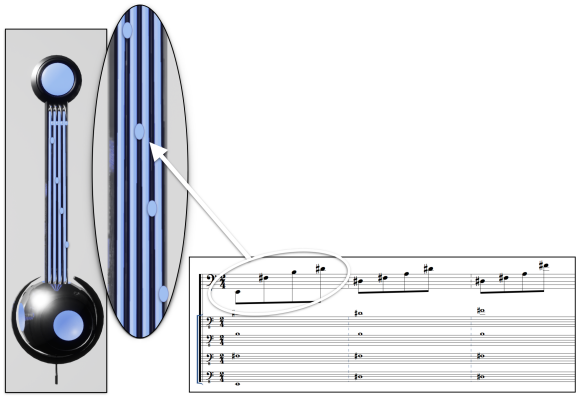


Fig. 4. MIDI score data sets score markers on the neck of the Coretet cello.

two movements of *Trois Machins de la Grâce Aimante* improvisation (Movement I) and a graphic reference score (Movement II) are used to convey individual and ensemble instructions out of real-time. However for Movement III, Coretet displays notes from a composed musical score in real-time as glowing blue pitch location indicators along the instrument's neck (see Figure 4). Scores are synchronized from the server to each of the clients and the system ingests individual MIDI tracks exported from a parent score using music notation software such as Finale.

4 Performance Practices

Coretet was designed primarily as a musical instrument for live concert performance and as such initial validation and testing of the system's technological and functional design decisions has taken place during ensemble rehearsal and live concert performances. The first musical work written for Coretet was *Trois Machins de la Grâce Aimante*, a three-movement string quartet composed to explore the sonic and interactive capabilities of the instruments and the collaborative and communicative aspects of the virtual networked environment. *Trois Machins de la Grâce Aimante* was premiered in Graz, Austria on September 27, 2019 at the Institut für Elektronische Musik und Akustik (IEM) Cube. To date the piece has had additional performances in Mexico City by a second ensemble, and again in Graz at the Mumuth's György-Ligeti-Saal.

4.1 *Trois Machins de la Grâce Aimante*

While *Trois Machins de la Grâce Aimante* is a composition intended to explore Twenty-First century technological and musical paradigms, it is at its heart a string quartet fundamentally descended from a tradition that spans back to the 18th century. As such, the work primarily explores timbral material based around

Fig. 5. Performers for *Trois Machins de la Grâce Aimante* sit in a semi-circle in traditional string quartet order [l to r: violin I, violin II, viola, cello] but face outward towards their tracking towers. In virtual space, each performer retains the same position but is facing inward to see one another's communicative gesture.

the sounds of bowed strings, in this case realized using physically modeled bowed strings, as well as ensemble communication and cooperation between the four performers. The composition takes the form of three distinct movements, each exploring different capabilities of the instrument itself and requiring different forms of communication and collaboration between the four performers [7].

4.2 Ensemble Position and Communications

To date, each performance of *Trois Machins de la Grâce Aimante* has taken place in a venue and in a format in keeping with Western concert music performance. To accommodate four performers with four Oculus Rift head-mounted displays, four laptops and four sets of Oculus Touch controllers and tracking towers, the tightly spaced, inward facing seating arrangements traditionally used in string quartet performance had to be inverted, forcing the ensemble members to physically face away from one another (see Figure 5). In this manner, the Oculus Touch tracking towers could be spaced far enough from each performer to create a sizeable interaction zone within which the motions of their head-mounted displays and hand controllers could be tracked.

Within the shared network space, each performer however is seated in the same relative location but facing inward, allowing the ensemble to see one another's avatar head, bow and instrument. In this manner ensemble members can communicate using gestures traditionally associated with string quartet performance practice, allowing for synchronized timing and visual cues for intensity. In the second movement of *Trois Machins de la Grâce Aimante*, the ensemble is required to conduct the playing of unison chords to one another using their bows, communicating tempo, note duration and intensity visually.

5 Discussion and Evaluation

Musical instruments offer a unique challenge for simulation in virtual reality. Traditional musical instrumental performance practices have been refined over hundreds of years by millions of end-users ranging in skill level from absolute beginner to expert virtuoso. Any attempts to simulate existing musical instruments - or even to create new virtual instruments that essentially inherit performance practice from existing instruments - are instantly and widely critiqued by users who share an intimate and personal connection with the parent model.

The physicality of traditional instruments is an inherent issue with current virtual reality systems. Without haptic feedback, a significant and functional core mechanic of musical instrument performance practice is removed, requiring novel and creative work-arounds to retain the missing functionalities in a manner that is essentially transparent to the end-user. And while today's generation of virtual reality displays and hand-tracking controllers offer an ease of use and widespread availability unprecedented in the history of VR systems, the control systems offered - ranging from bulky baton-controllers and joystick-like grips to infrared tracking of a performers hands - offer a completely different experience from a musician's learned behaviors of hand and finger interaction.

Similarly troubling are the low thresholds for latency required for real-time musical performance. Musicians are notoriously attentive to imposed latencies in musical systems, with thresholds of acceptable real-time latency residing well below 100 ms [1]. The graphics rendering and network requirements of a project like Coretet requires significant computing power across multiple machines, requiring a significant investment in computer and graphics card hardware.

From the time of its inception, Coretet was intended to exist as an instrument firmly descended from traditional ensemble instrumental performance practices with a goal of exploring how virtual implementations of musical instruments could leverage learned expert behaviors of highly skilled musicians. While Coretet's grounding in traditional performance practice has helped focus and constrain potential sonic and gestural exploration, as with any musical work composed for a novel musical interface, instrument or system there was a significant up-front investment of time and resources necessary to both design, create and then explore the physical and virtual affordances made available by Coretet. And as a vehicle for research and further exploration of virtual instrument design and performance, *Trois Machins de la Grâce Aimante* is currently being used to explore issues relating to the role of expert/virtuosic performance in game and instrument design, the perception of mixed and virtual reality performance by audience and performers alike, and the role and successes of communication between performers within rendered and virtual environments.

5.1 Expert Performance

While great success has been had by computer music instruments designed for non-musicians or casual musicians [6], Coretet is inherently designed for trained bowed string musicians, with a goal of leveraging learned performance practices

to create a virtual and futuristic version of a traditional instrument. In this context trained musicians share many similarities with expert game-players [12, 20] such as sensitivity to perceived latencies, especially across networks, and an ability and interest in exploring subtle and extended techniques and capabilities of the system. And while each recent performance (and the composition of *Trois Machins de la Grâce Aimante* itself) has stressed ensemble articulation and voicing over individual virtuosic gesture, in post-concert evaluations, within the context of the GAPPP research project, virtuosity in *Trois Machins de la Grâce Aimante* and Coretet was rated significantly higher than in other GAPPP projects evaluated similarly.

5.2 Evaluation by Audience and Performers

In conjunction with the GAPPP project, the premiere performance of *Trois Machins de la Grâce Aimante* featured a post-concert written and oral evaluation of the work by members of the audience, as well as real-time attention tracking during the performance through the use of networked tablets with touch interfaces distributed to individual audience members [17]. While analysis of results from this evaluation are still preliminary, one early reported finding was a significant perception of the virtuosity inherent in performance, a characteristic less commonly associated with other projects as presented and analyzed through the GAPPP metrics. Subjects also reported a desire to view the performance in VR using their own head-mounted displays, a suggestion in keeping with the intent of increasing their own sense of immersion.

5.3 Communication

As elements of musical communication and collaboration are traditional components of ensemble performance and string quartet performance practice in particular, significant consideration went into creating Coretet's network and replication layer to allow performers to see one another. Conversations with performers both during the initial workshop and premiere of the work in Graz, Austria as well as with a second set of performers during the 2019 Ecos Urbanos festival in Mexico City, Mexico showed that performers felt able to see and communicate to a certain extent within the virtual space using methods based in their existent performance practice. Suggestions for greater communication included lessening the perceived distance between virtual performers to make subtle gestures more perceivable as well as rendering more virtual body parts such as arms and hands. During these initial performances a software bug made performers' heads not visible to the rest of the ensemble, which was reported as greatly lessening performers' abilities to non-verbally communicate with one another.

6 Conclusions

The design and development of Coretet was intended as an exploration of VR systems as possible conduits for musical expression. Coretet was envisioned as an instrument directly descended from traditional stringed instrument design as perfected by luthiers throughout the ages. In that light, Coretet was designed first as an instrument for trained musicians, with a future intent of creating levels of abstraction to allow less trained musicians to use the application.

Successful performances using Coretet of *Trois Machins de la Grâce Aimante* suggest that the modes of networked virtual reality instrumental performance afforded by the Coretet instrument are a viable form of musical performance and that the introduction of virtual reality systems within a real-time musical concert situation - while novel - can be perceived as being fundamentally similar to that of traditional musical instrumental performance. Expert users have indicated that performing with the Coretet instrument in *Trois Machins de la Grâce Aimante* has significant commonality with traditional string quartet performance practice and seems to leverage learned skills on stringed instruments. And audience members queried after watching a performance specifically noted the quality, collaboration, virtuosity and creative nature of the project.

Ongoing development of Coretet is currently focused on improving the user experience, bug fixing and implementing interface layers sufficient to submit the application to online software stores available for download by end users.

Acknowledgments. Support for Coretet and the composition of *Trois Machins de la Grâce Aimante* was generously provided by a GAPPP residency and commission from the Institut für Elektronische Musik und Akustik (IEM) in Graz, Austria. Hardware support was provided by an Nvidia hardware grant and additional funding and support was provided by Rensselaer Polytechnic Institute.

References

1. Bartlette, C., Headlam, D., Bocko, M., Velikic, G. Effect of Network Latency on Interactive Musical Performance, Music Perception; Berkeley Vol. 24, Iss. 1, September: 49–62. (2006)
2. Berdahl, E., Backer, S. and Smith, J., III.: If I Had A Hammer: Design and Theory of an Electromagnetically Prepared Piano. In: Proceedings of the International Computer Music Conference, Barcelona, Spain, 59, September (2005)
3. Cook, P. and Scavone, G.: The Synthesis Toolkit (STK). In: Proceedings of the International Computer Music Conference, Beijing, China (1999)
4. Dolby, T.: The Speed of Sound: Breaking the Barriers Between Music and Technology: A Memoir. p. 166, Flat Iron Books, MacMillan, New York (2016)
5. Hamilton, R.: Mediated Musical Interactions in Virtual Environments. In: S. Holland, K. Wilkie, T. Mudd, A. McPhearson, M. Wanderley (Eds.) New Directions in Music and Human-Computer Interaction, Springer, Heidelberg (2019)
6. Hamilton, R., J. Smith, G. Wang.: Social Composition: Musical Data Systems for Expressive Mobile Music. Leonardo Music Journal, Vol. 21, pp. 57–64 (2011)

7. Hamilton, R.: *Trois Machins de la Grâce Aimante*: A Virtual Reality String Quartet. In: Proceedings of the International Computer Music Conference, New York City, USA (2019)
8. Hamilton, R.: Maps and Legends: Designing FPS-based Interfaces for Multi-User Composition, Improvisation and Interactive Performance. In: Computer Music Modeling and Retrieval, LNCS, vol. 4969, pp. 478–486. Springer, Heidelberg (2007)
9. Hamilton, R., Platz, C.: Gesture-based Collaborative Virtual Reality Performance in Carillon, In: Proceedings of the International Computer Music Association Conference, Utrecht, Netherlands (2016)
10. Hemingway, T. Turn It Up to Eleven: A Study of Guitar Hero and Rock Band: Why People Play Them and How Marketers Can Use This Information. Masters Thesis, Brigham Young University (2010)
11. Leonard, J. and Cadoz, C.: Physical Modelling Concepts for a Collection of Multisensory Virtual Musical Instruments. In: Proceedings of the 2015 International Conference on New Interfaces for Musical Expression, Baton Rouge, USA (2015)
12. Lindsted, J., Gray, W.: Distinguishing experts from novices by the Mind's Hand and Mind's Eye. Cognitive Psychology, Vol. 109, March, pp. 1–25 (2019)
13. Lüneburg, B.: Between Art and Game: Performance Practice in the Gamified Audiovisual Artworks of GAPPP. The Computer Games Journal 7:243–260, Springer Science (2018)
14. Mäki-Patola, T., Kanerva, A., Laitinen, T. and Takala, T.: Experiments with Virtual Reality Instruments. In: Proceedings of the 2005 International Conference on New Interfaces for Musical Expression, Vancouver, B.C., Canada (2005)
15. McPherson, A. and Kim, Y.: Augmenting the Acoustic Piano with Electromagnetic String Actuation and Continuous Key Position Sensing. Proceedings of the 2010 International Conference on New Interfaces for Musical Expression, Sydney, Australia (2010)
16. Overholt, D., Berdahl, E., Hamilton, R.: Advancements in Actuated Musical Instruments. Organised Sound, volume 16, issue 02, pp. 154–165. (2011)
17. Pirchner, A.: IRMA (Interactive Real-time Measurement of Attention). A new method investigating performances of audiovisual computer music. In: Proceedings of the International Computer Music Conference, New York City, USA (2019)
18. Puckette, M.: Pure Data. In: Proceedings of the International Computer Music Conference. San Francisco, 1996, pp. 269–27 (1996)
19. Raaen, K., Kjellmo, I.: Measuring Latency in Virtual Reality Systems. 14th International Conference on Entertainment Computing (ICEC), Sep 2015, Trondheim, Norway. Lecture Notes in Computer Science, LNCS, vol. 9353, pp.457–462. Springer, Heidelberg (2015)
20. Salice, A., Høffding, S., Gallagher, S.: Putting Plural Self-Awareness into Practice: The Phenomenology of Expert Musicianship. Topoi (2017)
21. Serafin, S., Erkut, C., Kojs, J., Nilsson, N., and Nordahl, R.: Virtual Reality Musical Instruments: State of the Art, Design Principles, and Future Directions. Computer Music Journal, Vol 40:3, Fall, pp. 22–40, MIT Press, Cambridge, MA (2016)
22. Wang, G., Oh, J., Lieber, T.: Designing for the iPad: Magic Fiddle. In: Proceedings of the 2011 New Interfaces for Musical Expression Conference, Oslo, Norway (2011)
23. Wright, M. and A. Freed.: Open Sound Control: A New Protocol for Communicating with Sound Synthesizers. In: Proceedings of the International Computer Music Conference., Thessaloniki, Greece (1997)

Author Index

ISBN 979-10-97-498-01-6
Les éditions de PRISM